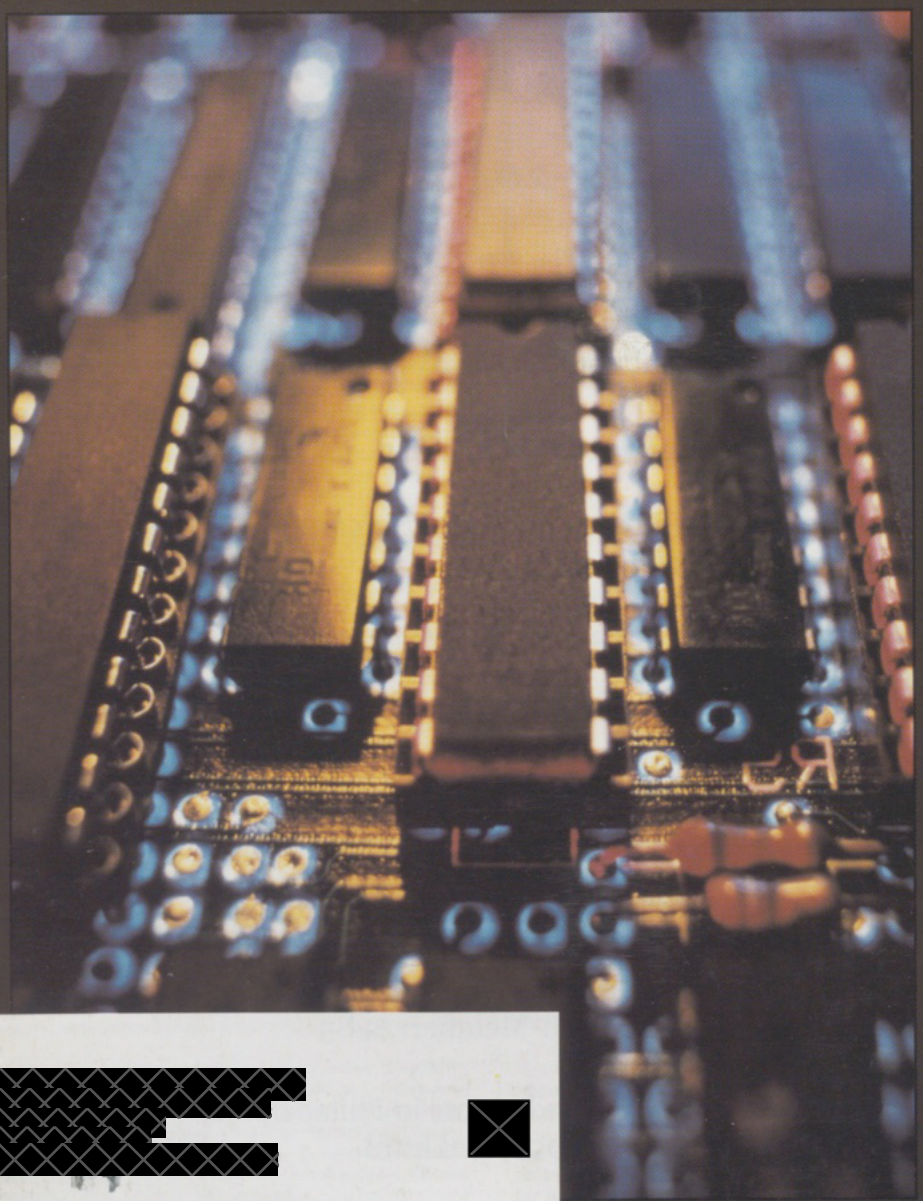
Information Technology AND Libraries

ISSN 0730-9295

Volume 21 Number 4

DECEMBER 2002



IN THIS ISSUE

LITA

Workshop

at the

2002 ALA Midwinter Meeting

in Philadelphia

XML and Libraries

January 24, 2003

9 A.M.–5 P.M.

Pennsylvania Convention Center–204 A

Register now at www.lita.org/mw2003/xml.html.
(Registration for Midwinter is not required to attend the workshop.)

Recent workshop participants had this to say about **XML and Libraries**:

"The 'nuts and bolts' of XML were very clearly explained . . ."

"I enjoyed the style of the presentation and the thoughtfulness with which questions were answered."

"Ron Gilmour was a fantastic presenter and his knowledge and coverage of the topic was fantastic."

Registration Fees: LITA Member: $165; ALA Member: $210 (includes LITA membership); Non-Member: $285

For information on other LITA meeting and events in Philadelphia, please visit www.lita.org/ac2002/xml.html

Information Technology and Libraries (ISSN 0730-9295) is published quarterly in March, June, September, and December by the American Library Association, 50 E. Huron St., Chicago, IL 60611. It is the official publication of the Library and Information Technology Association, a division of the American Library Association. SUBSCRIPTIONS: Subscription price, $30, is included in membership dues. Nonmembers may subscribe for $55 per year in the U.S.; $60 in Canada, Mexico, Spain, and other PUAS countries; $65 in other foreign countries. Send subscription requests to *Information Technology and Libraries,* Subscription Department, American Library Association, 50 E. Huron St., Chicago, IL 60611; 800-545-2433; fax: (312) 944-2641; subscription@ala.org. Single copies, $20. Periodical-class postage paid at Chicago, Illinois, and at additional mailing offices. POSTMASTER: Send address changes to *Information Technology and Libraries,* 50 E. Huron St., Chicago, IL 60611.

Information Technology and Libraries publishes material related to all aspects of libraries and information technology, including digital libraries, metadata, authorization and authentication, electronic journals and electronic publishing, telecommunications, distributed systems and networks, computer security and intellectual property rights, technical standards, geographic information systems, desktop applications, online catalogs and bibliographic systems, optical information systems, software engineering, universal access to technology, futuristic forecasting, library consortia, vendor relations, and technology and the arts. *ITAL* welcomes unsolicited manuscripts. Submissions should follow the guidelines stated under "Instructions to Authors" on page 92 of the June 2002 issue or on the LITA Web site at www.lita.org/ital/infoauth.htm. Manuscripts of feature articles, communications, and tutorials should be addressed to Dan Marmion, 221 Hesburgh Library, The University of Notre Dame, Notre Dame, IN 46556-5629. Books for review should be addressed to Tom Zillner, WILS, 464 Memorial Library, 728 State St., Madison, WI 53706. Software for review should be addressed to Andy Boze, Hesburgh Library, The University of Notre Dame, Notre Dame, IN 46556-5629.

As a matter of policy, *ITAL,* as the scholarly organ of LITA, does not review LITA publications.

PRODUCTION: ALA Production Services (Troy D. Linker, Kevin Heubusch; Ellie Barta-Moran, Angela Hanshaw, Kristen McKulski, and Karen Sheets), American Library Association, 50 E. Huron St., Chicago, IL 60611.

Publication of material in *Information Technology and Libraries* does not constitute official endorsement by LITA or the ALA.

Abstracted in *Computer & Information Systems, Computing Reviews, Information Science Abstracts, Library & Information Science Abstracts, Referativnyi Zhurnal, Nauchnaya i Tekhnicheskaya Informatsiya, Otdyelnyi Vypusk,* and *Science Abstracts Publications.* Indexed in *CompuMath Citation Index, Computer Contents, Computer Literature Index, Current Contents/Health Services Administration, Current Contents/Social Behavioral Sciences, Current Index to Journals in Education, Education, Library Literature, Magazine Index, NewSearch,* and *Social Sciences Citation Index.* Microfilm copies available to subscribers from Proquest, Ann Arbor, Michigan.

The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences—Permanence of Paper for Printed Library Materials, ANSI Z39.48-1992.∞

Information Technology AND Libraries

Volume 21, Number 4 December 2002 ISSN 0730-9295

Abstracts and contents pages of recent issues of *ITAL* can be found on the LITA Web site at www.lita.org/ital/index.htm.

cover design Kevin Heubusch ● **interior design** Dianne M. Rooney

Editorial: Looking at the Publishing Process

Dan Marmion

I recently responded to some questions posed by Rachel Singer Gordon, who is gathering information for a book she will publish in 2004, and I thought they might be of interest to *ITAL* readers and potential authors.

Q. What should authors expect after submitting a manuscript to ITAL?

A. There are always exceptions, but in general here is what happens. I first give it a cursory examination to see whether there is an obvious reason to reject it, e.g., it is clearly out of the scope of *ITAL*. Then it goes into the peer-review process, which should only take three to four weeks but sometimes is longer, especially if I decide I want a second opinion (which doesn't happen very often). Depending upon the reviewer's recommendation (see next question), the manuscript could be accepted or rejected or, most likely, could require some rewriting, in which case there would be an additional time interval. Once we are ready to accept a manuscript, I ask the author to send me an electronic version of the final document.

The production process involves a period during which the manuscripts are copyedited. Then we contact authors with questions identified by the copyeditor in the first pass through. These could be any number of things, such as a reference that doesn't look correct, a sentence or paragraph is not clear as to what the author means, is a particular edit acceptable to the author, and so on. Our managing editor, Judith Carter, works with the authors to resolve these questions, then communicates the results back to ALA Production Services, where they are incorporated into the emerging issue. A second pass always generates a few more questions that also need to be resolved. Sometimes, although rarely, an article needs a third pass before we can declare the issue ready to go to the printer.

Q. Please explain how the review process works at ITAL, *and what reviewers examine when looking at a manuscript. Do you employ a regular board of reviewers, or send manuscripts to subject experts as needed? Are all sections of the journal refereed?*

A. We follow a typical double-blind review process, meaning that the author doesn't know who reviews the manuscript and the reviewer doesn't know who wrote it.

Reviewers are asked to consider several things as they assess a manuscript's potential for publication: Is the topic within the scope of *ITAL?* Is it meaningful and relevant to *ITAL* readers? Does it offer something to the literature? Is it timely? Is it organized well? Does it have a point? Are the citations complete and accurate? Does the manuscript accomplish whatever the author set out to do? Is there something more that needs to be done to make it a more solid contribution?

The reviewer makes a recommendation to me, which is one of the following: (1) accept it and publish as is, or with minor editorial changes; (2) it's almost there but requires some rewriting to make it a solid publishable contribution; (3) it shows some potential, but requires major rewriting and should be reviewed again after a revised draft has be submitted; or (4) it does not warrant further consideration for *ITAL*. In all but the first instance, the reviewer should provide specific feedback regarding the manuscript's shortcomings. I pass along to the author any feedback or criticism given to me by the reviewer, paraphrasing it as necessary to preserve the anonymity of the reviewer.

The *ITAL* Editorial Board does the majority of reviews; that is their primary duty, and I am fortunate to have a board made up of experienced persons with much expertise in many areas of information technology and libraries. It is not uncommon, however, for me to decide that an expert in the field would best review a particular manuscript.

The only sections of *ITAL* not refereed are the editorial page and the book review and software review columns.

Q. Approximately what percent of article submissions to ITAL *are accepted? How often are works sent back to the author for further revision, and what is usually entailed in the revision process?*

A. Of all the manuscripts submitted to *ITAL*, approximately 50 to 60 percent are ultimately accepted for publication, but probably at least 90 percent of those required further revision by the author. If the reviewer has indicated that the manuscript is pretty close to publishable, I'll just tell the author what more we request in the way of a rewrite and ask them to make those revisions and send me the revised manuscript with a cover letter indicating what they did to satisfy the request. I verify that they did so, and we're done.

In those cases where the reviewer has indicated a major rewrite is necessary and that the revised manuscript should be reviewed again, I convey that information and any other suggestions to the author. I ask them to indicate to me whether they plan to do that and submit a revised manuscript. Some I never hear from again; others respond in a negative tone. The majority thank me for the suggestions and promise to submit a revised piece based upon the criticism offered. Those persons almost always follow through with a much better manuscript that the one they originally submitted. It may still take one or two subsequent revisions to get it into top-notch shape, but the end result is a much better work.

Dan Marmion (dmarmion@nd.edu) is Associate Director for Information Systems and Digital Access at the University of Notre Dame Libraries, Notre Dame, Indiana.

Network CDs... Without CDs

Mirror 1 to 512 CDs/DVDs onto a massive hard disc and share the content among all clients.

The SuperCD server lets everyone share the data concurrently with fast, web-browser accessibility. It's the effortless and affordable way to access reference material, archived records, and computer application programs

Unnecessary telecom costs are eliminated. Content control hassles are minimized, as you pre-select the library. Hacker mischief is thwarted, since SuperCD is a read-only server. Extra CD-ROM drives and multiple CD copies are eliminated for easier administration. And technical support calls are a rarity.

Relegate anachronistic CD juke boxes to history. Join the thousands of satisfied SuperCD users in libraries, schools, and offices today.

SuperCD

CD/DVD/CDR Server

- **1 to 512 CDs or DVDs**
- **NT, Netware & Mac**
- **Web Browser Accesible**
- **Auto CD License Monitoring**
- **Record Custom CDs**
- **66x Transfer Rates**
- **Free 30 Day Trial**
- **One Year Warranty**

AXONIX®

IT Solutions Since 1983

800.866.9797

www.axonix.com

An Empirical Analysis of Web Catalog User Experiences

Dennis Halcoussis, Aniko L. Halverson, Anton D. Lowenberg, and Susan Lowenberg

Data from an observation study of a Web catalog in a small private arts college library are used to analyze the determinants of user success and satisfaction. Multiple regression models are estimated to identify the most important causative factors determining catalog user success in finding information, user attitudes to catalog organization, and user ability to navigate the catalog. It is found that subject-search users are more likely to assign a low score to catalog organization and to encounter difficulty navigating the catalog than users of known item and other search methods. These findings accord with the extensive literature on the problems associated with subject searching. Also, it is found that the more time spent searching and the larger the number of search results, the more likely it is that the user would report difficulty navigating the catalog. A significant result is that although the user's perception of success or failure of the search is the most important factor determining both the user's evaluation of the catalog organization and the navigability of the catalog, the success or failure of the search itself is not explained by any other variables included in the model. This exogeneity of search success has important implications for library instruction because it suggests that a user's perception of success is dependent on the expectations the user brings to the search rather than specific features of the catalog design.

Use studies of online catalogs reveal several empirical regularities, namely that users adapt to the tools provided and typically prefer online systems, and that subject retrieval capabilities are inadequate.[1] According to Lewis, users "do not understand the complexities of bibliographic structures in any form" and "inconsistencies in cataloging practices cause confusion" for users.[2] Moreover, use studies consistently indicate that the information users bring to catalog searches is often incomplete, and that users are normally more successful in conducting known item searches than subject searches.

That subject searching, in particular, is troublesome for users is well documented in the literature.[3] Of all the various forms of online catalog searches, users generally report the most difficulty with subject searches.[4] A seminal nationwide study of online catalog use sponsored by the Council on Library Resources found that subject searching was most likely to prove problematic for users, with 43 percent of users reporting difficulties with this search method.[5] The main reasons for these difficulties cited in the literature are misspelling of search terms and lack of knowledge of Library of Congress subject headings.[6] Database growth, which produces excessive numbers of results, also causes problems for users.[7]

Given these findings, is there anything an individual library can do to improve its online catalog and thus enhance the quality of service to users? To address this question, we conducted an observation and questionnaire study at a small arts college, the results of which are reported here. The purpose of our study is to identify the determinants of user success with, and attitudes toward, a new Web-based online catalog. We analyze our findings using multiple regression analysis, a departure from most of the existing literature. The multiple regression approach enables us to investigate how various user attributes and search types, as well as characteristics of searches recorded in the observation study, affect search outcomes. Our results will provide guidelines for future improvements in bibliographic instruction and user training.

Literature Survey

Numerous studies of the use of online catalogs have been conducted since they were introduced in the early 1980s. The first generation of online catalogs had minimal capabilities, being little more than library circulation systems made available to general users, with the result that many of the earlier studies have limited applicability to today's catalogs with their advanced subject and keyword searching functions.[8] Consequently, we focus on those studies conducted since the late 1980s, which can be classified into the following main categories: questionnaires and surveys, transaction monitoring, observation, and experiments.

Questionnaires and surveys and transaction log analysis have been by far the most popular approaches used in the study of online catalogs. A few researchers have utilized a combination of two or more techniques.

Dennis Halcoussis (dennis.halcoussis@csun.edu) is Professor, Department of Economics, California State University, Northridge; **Aniko L. Halverson** (coco@calarts.edu) is Reference Coordinator and Instruction Librarian, Division of Library and Information Resources, California Institute of the Arts, Valencia; **Anton D. Lowenberg** (anton.lowenberg@csun.edu) is Professor, Department of Economics, California State University, Northridge; and **Susan Lowenberg** (susan@calarts.edu) is Associate Dean, Division of Library and Information Resources, California Institute of the Arts, Valencia.

However, many existing studies are plagued by methodological problems.[9] According to Seymour, "a general failure in the use of statistical methods is apparent. Poor sampling, generalization from small samples, 'torturing the data' . . . and application of inappropriate tests . . . abound."[10]

Only a handful of recent questionnaires and surveys employ statistically valid sample sizes and appropriate statistical tests. For example, Ensor studies user characteristics of keyword searching.[11] Hsieh-Yee utilizes a mailed questionnaire and chi-squared tests to examine the relationship between user characteristics and information-seeking behavior.[12] Park implements a four-part questionnaire, including skills tests and user evaluations, to compare five selected online catalog systems.[13]

Transaction monitoring of online catalog logs is a fairly common approach in the literature. Sinnott uses transaction log monitoring to examine the frequency of no direct hits and zero hits in author searches.[14] She compares the percentage of no direct hits in 1993 with that ten years earlier and finds a decline in this percentage over the decade. A statistically valid sample size of 220 cases and the implementation of a t-test for proportions make her conclusions credible. Blecic et al. report a longitudinal study of four sets of online catalog transaction logs over a four-year period to test whether a series of screen changes helped users improve search behavior.[15] Again, a large sample size for each set of logs and use of binomial distribution techniques lend credibility to their conclusions.

In one of the few studies to utilize multiple regression analysis, Larson estimates the effects of a linear time trend and seasonal factors (summer and holiday breaks) on the frequency of subject searches and known item searches over six years of transaction monitoring data for the University of California's MELVYL online catalog.[16] His findings indicate a persistent decline in the percentage of subject searches over the data collection period and a corresponding increase in the percentage of known item, especially title keyword, searches.[17] He also finds a dramatic drop in subject searching and rise in title keyword, author, and other known item searching during summer and holiday breaks, when most library users are likely to be those with the greatest experience and familiarity with the catalog.[18] In addition, he reports a statistically significant positive correlation between the failure rate for all types of searches and the percentage of subject searching. These results taken together suggest that users, over time, learned to substitute title and other known item searches for subject searches in response to the frequent failure of subject searches.

Some studies have used a combination of questionnaires and transaction log analysis. Wiberley, Daugherty, and Danowski measure the statistical correlation between user persistence in scanning retrieved search results and the prior search overload of the user.[19] They find that overloaded users appear to have lower thresholds for optimal information processing. Users with a lower capacity for processing typically display perceptions of overload more readily, so that for these individuals relatively low numbers of postings can trigger overload. Hildreth uses cross-tabulation analyses and a chi-squared test to study keyword searching in an online catalog, concluding that although users search more often by keyword than by any other method, most keyword searches fail and "a majority of users do not understand how the system processes their keyword searches."[20] Dimitroff utilizes both personal interviews and transaction log analysis to examine the effects of users' mental models of a bibliographic retrieval system.[21]

In an interesting experiment, Drabenstott and Weller compare the use of search trees relative to other subject searching approaches.[22] They examine 528 search administrations to identify those factors that result in subject search failure. Causes of search failure are shown to include lack of user perseverance in displaying retrieved titles, queries that are not indicative of user interests, conflicting relevance assessments, and design problems with particular subject-searching approaches. Search trees are found to improve user search performance, and users generally prefer search trees over alternative systems on the grounds of superior search outcomes, ease of use, efficiency, clarity, and other criteria. Although the sample size used in this experiment is certainly adequate, the authors do not implement sophisticated statistical techniques to analyze their data beyond simple t-tests.

Observation studies are not heavily represented in the literature, probably due to their relatively high cost. Belkin et al. employ a variety of assessment methods, including an observation study of two hundred users, to document the relationship between user characteristics and goals on the one hand and search behavior on the other.[23] Their observation of user information behavior, both on the catalog and subsequently in the library as a whole, leads them to conclude that online systems might be improved substantially by including tables of contents in item records, offering users a map from the terminal to the location of the material identified, and providing users, especially of public libraries, records of their prior borrowing activities. Hert and Nilan combine interviews and observations of ninety-three subjects in order to measure the degree of match between a catalog user's desired action and the actual system command entered by the user, i.e., the extent to which a command issued to the system is appropriate for the action intended by the user.[24] Results are presented descriptively, along with data on the insights or knowledge that users draw upon in formulating their searches. The authors find that users who rely on the catalog system for insight in initiating a search generally fare better than users who rely on their own knowledge, although

the degree of match for system insight is still quite low, and the system performs poorly in helping users revise searches.

Cherry reports a study in which observers recorded online catalog use and then asked users to fill out a questionnaire about their background.[25] However, the validity of the results is marred by a small sample size (only ninety-four usable observations were collected), and the analysis of the findings does not extend beyond simple descriptive statistics. Thorne and Whitlach likewise use a combination of observations and a questionnaire to examine patron success on an online catalog.[26] Their study, also limited to descriptive statistics, again is hampered by a small sample size of only ninety-three observations. Anderson conducts an observation study of online search techniques of fifty university students, recording the type of search, use of library personnel and online help screens, exposure to library instruction, and length of time at the terminal.[27]

Overall, most of the existing literature on online catalog use suffers from small sample sizes, unsophisticated statistical tests of significance, and simplistic, largely descriptive, presentations of findings. In the present study we propose to remedy these shortcomings by undertaking a multiple regression analysis of the determinants of search outcomes.

Background

The California Institute of the Arts (CalArts) is a private institution of higher education created specifically for students of both the visual and performing arts, emphasizing the contemporary forms of these arts. CalArts grants BFA and MFA degrees in art, dance, film/video, music, and theater as well as an MFA in writing. The CalArts Library first instituted an online catalog, CALIS, in spring 1993 using the Data Research Associates (DRA) Information Gateway module. In summer 1998 CALIS was upgraded to DRAWeb2, a Web-based online catalog. One feature of the DRAWeb2 catalog is the ability of the library to tailor HTML pages to its needs. The CalArts Library's implementation of the DRAWeb2 catalog has been highly customized, with modifications made to the navigational structure, page layout, and record display. When there are major changes to a vital library service, such as an online catalog, there are often vocal members of the community who dislike the new interface, preferring the old one. The library wanted to determine if this feeling was widespread among the CalArts community and, if so, what changes could be made to improve either the Web catalog or library instruction.

Method

Two survey instruments were developed. The first was an observation form that was used to record the actions of the catalog user, including the search term entered, the search screen used, the type of search selected, and the number of results. The observers were two professional librarians, three staff members, and several student workers. Observers were instructed to approach users of the Web catalog and ask them to participate in the study. Users who had already participated were not surveyed again, so that the data would contain only unique observations. Observers were also instructed to take notes as they watched users search the catalog, but not to interact with the users or to answer any questions until after the users had completed the questionnaire. See appendix A for the observation study data collection sheet.

The second survey instrument was a questionnaire that asked users how long they had been at CalArts, their experience using the catalog, their academic status at CalArts, whether English was their native language, and whether they had found what they were looking for in their catalog search. In addition users were asked to rank their responses to the following two statements on a five-point scale: "The organization of the CalArts Library catalog is effective for finding information" (1 = not effective; 5 = very effective) and "When I was looking for information, I felt lost" (1 = felt lost; 5 = felt I knew where I was).[28] See appendix B for the questionnaire.

There were 1,230 registered students at CalArts during the 1999–2000 academic year and 372 full time-equivalent employees (faculty and staff), giving a total CalArts population of 1,602. A sample size of at least 216 is necessary to obtain statistically significant results at a 10 percent error level in a two-tailed test, with a confidence interval of plus or minus 5 percent.[29] The observation study was conducted from December 1999 through April 2000. A total of 223 survey responses were collected, thus providing a sufficiently large sample to produce statistically valid results.

The observation data were recorded in a spreadsheet. If multiple searches were performed, only data from the last search were recorded, on the grounds that the last search most accurately reflects the user's perception of search success or failure. For each search, the observer recorded the type of search (author, title, subject, keyword, author keyword, title keyword, or subject keyword), the length of time for the entire session, and the number of results. Personal information about the user collected with the questionnaire was also recorded. The data were analyzed using SPSS, a software program that handles a variety of estimation methods.

Empirical Results

The survey results showed that the majority of Web catalog users were successful in finding the information they sought, satisfied with the organization of the catalog, and able to find their way around in the catalog. Specifically, 70 percent of users successfully found what they were looking for in their Web catalog search. Fifty-three percent of users felt that the organization of the Web catalog was effective or very effective for finding information, assigning the catalog organization a rating of 4 or 5 on the five-point scale (1 = not effective; 5 = very effective). Only 14 percent of users felt that the catalog organization was not effective, giving a rating of 1 or 2.

Sixty-five percent of users reported that they knew where they were when using the Web catalog, awarding the catalog a rating of 4 or 5 on the five-point scale (1 = felt lost; 5 = felt I knew where I was). Thirteen percent of catalog users indicated that they felt somewhat lost, giving the catalog a rating of 1 or 2.

As mentioned above, given a sample size of 223, these results are accurate to within plus or minus 5 percent, with only a 10 percent chance that the actual percentages in the overall library patron population fall outside this range.

The data were also analyzed using multiple regression analysis in order to identify possible causal relationships between the variables recorded in the observation study and the attitudes and success of users. The regression results are presented in tables 1 through 4. Each coefficient estimate reported in the tables gives the average change in the dependent variable for a one-unit change in the independent variable, holding the other independent variables constant. For example, in table 1 the dependent variable is ORGANIZATION, users' rankings of the organization of the catalog. One of the independent variables is AUTHOR, which indicates an author search. The coefficient estimate for AUTHOR is 1.310. This means that, on average, if a user implements an author search, he or she will give the organization of the catalog a rating that is 1.310 points higher on a five-point scale than someone who did a subject search, holding all other variables in the model constant. Subject searches are omitted from the regression model because these are treated as the default or control category.[30] If the sign of a coefficient estimate is negative, then there is a negative relationship between that variable and the dependent variable. For example, the coefficient estimate for the MUSIC variable in table 1 is -1.034, meaning that users from the school of music, on average, assign the catalog organization a rating that is 1.034 points lower than users from the school of art (the control category), holding all other variables in the model constant.

Estimated coefficients should only be interpreted in the way described above if they are statistically significant at a 10 percent error level or better. If the error level, also referred to as the significance level, is higher than 10 percent, this means that there is a reasonable chance that the true value of the coefficient is zero. Significance levels are provided in the fourth column of each table. In order for a coefficient estimate to be significant at a 10 percent error level or better, the entry in the fourth column must be less than or equal to 0.10.

Table 1. Ordinal Regression: Dependent Variable Is ORGANIZATION

Dependent variable	Coefficient estimate	Standard error	Significance level
SUCCESS	1.192	.349	.001
AUTHOR	1.310	.502	.009
TITLE	.919	.493	.062
KEYWORD	.633	.447	.157
TITLE KEYWORD	-1.711	1.441	.235
TIME	3.394E-02	.033	.299
RESULTS	-2.593E-04	3.1E-04	.398
YEARS	.195	.171	.252
EXPERIENCE	-6.875E-02	.165	.676
ESL	.430	.416	.301
MFA	-.789	.360	.028
FACULTY	-.343	.739	.643
STAFF	1.332	1.428	.351
DANCE	2.208	1.628	.175
MUSIC	-1.034	.422	.014
THEATER	-.898	.439	.041
FILM	-.377	.453	.405

R^2=.23

The regression model in table 1 estimates the effects of a number of causative factors on a user's perception of catalog organization in terms of its effectiveness in finding information. The dependent variable is ORGANIZATION, the user's ranking of the catalog organization, which, as indicated above, is measured on a five-point scale ranging from 1 (not effective) to 5 (very effective). Because this variable can take only five discrete values, an ordinal regression equation is estimated.[31] Causative factors included in the regression model are the user's perception of success or failure of the search; type of search; duration of the search; number of results yielded by the search; and several personal characteristics of the user,

Table 2. Ordinal Regression: Dependent Variable Is LOST

Dependent variable	Coefficient estimate	Standard error	Significance level
SUCCESS	1.144	.316	.000
AUTHOR	.718	.429	.094
TITLE	.469	.444	.292
KEYWORD	.688	.406	.090
TIME	-4.902E-02	.030	.104
RESULTS	-5.221E-04	2.8E-04	.066
EXPERIENCE	.186	.135	.169
ESL	-.501	.366	.171
DANCE	1.469	1.580	.352
MUSIC	-.617	.375	.100
THEATER	-.559	.391	.153

R^2=.17

Table 3. Ordinal Regression: Dependent Variable Is LOST

Dependent variable	Coefficient estimate	Standard error	Significance level
SUCCESS	1.099	.310	.000
AUTHOR	.894	.412	.030
TITLE	.531	.425	.211
KEYWORD	.805	.391	.039
TIME	-5.417E-02	.029	.064
RESULTS	-5.143E-04	2.8E-04	.071
MUSIC	-.550	.364	.131
THEATER	-.645	.382	.092

R^2=.15

such as his or her experience using the catalog, the number of years he or she had spent at CalArts, his or her academic status and school affiliation, and his or her English proficiency. Specifically, the independent variables in table 1 are defined as follows:

- SUCCESS = dummy variable taking on a value of one if the user indicated that he or she had been successful in finding what he or she was looking for, zero otherwise;

Table 4. Logit Regression: Dependent Variable Is SUCCESS

Dependent variable	Coefficient estimate	Standard error	Significance level
CONSTANT	.277	.829	.738
AUTHOR	.551	.576	.339
TITLE	.775	.543	.153
SUBJECT	.528	.578	.361
KNOWN ITEM	-.225	.566	.691
TIME	.057	.046	.216
RESULTS	.002	.002	.298
YEARS	-.232	.206	.260
EXPERIENCE	.105	.209	.614
ESL	-.326	.486	.502
MFA	.453	.430	.292
FACULTY	.145	.851	.865
DANCE	-1.641	1.616	.310
MUSIC	-.438	.516	.396
THEATER	1.193	.641	.063
FILM	-.729	.503	.147

R^2=.17

- AUTHOR = dummy variable taking on a value of one if the user implemented an author search, zero otherwise;
- TITLE = dummy variable taking on a value of one if the user implemented a title search, zero otherwise;
- KEYWORD = dummy variable taking on a value of one if the user implemented a keyword search, zero otherwise;
- TITLE KEYWORD = dummy variable taking on a value of one if the user implemented a title keyword search, zero otherwise;
- TIME = time spent on the search;
- RESULTS = number of results produced by the search;
- YEARS = number of years the user had been at CalArts;
- EXPERIENCE = number of times per week the user typically used the catalog;
- ESL = dummy variable taking on a value of one if English was the user's second language, zero otherwise;
- MFA = dummy variable taking on a value of one if the user was an MFA student, zero otherwise;

- FACULTY = dummy variable taking on a value of one if the user was a faculty member, zero otherwise;
- STAFF = dummy variable taking on a value of one if the user was a staff member, zero otherwise;
- DANCE = dummy variable taking on a value of one if the user was affiliated with the school of dance, zero otherwise;
- MUSIC = dummy variable taking on a value of one if the user was affiliated with the school of music, zero otherwise;
- THEATER = dummy variable taking on a value of one if the user was affiliated with the school of theater, zero otherwise; and
- FILM = dummy variable taking on a value of one if the user was affiliated with the school of film/video, zero otherwise;

In table 1, the coefficient of the SUCCESS variable has a positive sign and is statistically significant at the 1 percent error level. This means that users who rated their use of the catalog as successful were significantly more likely to be satisfied with the organization of the catalog than users who rated their search unsuccessful. The AUTHOR and TITLE coefficients are also positive and statistically significant, the first at the 1 percent significance level and the second at the 10 percent level. This means that users who implemented known item searches, such as author or title searches, were significantly more likely to award the organization of the catalog a high rating relative to users who implemented subject searches (the latter being the control group).

The MUSIC and THEATER coefficients in table 1 are both negative and statistically significant at the 5 percent level. This means that users from the music and theater schools were significantly less likely to assign the organization of the catalog a high rating than users from the art school (the control group). The only other statistically significant variable in table 1 is MFA, whose coefficient estimate is negative and significant at the 5 percent level, indicating that MFA students were significantly less likely to give the organization of the catalog a high rating compared to BFA students (the control group).

Overall, the independent variables included in table 1 explain 23 percent of the variation in organization ratings across users, as indicated by the R^2 statistic. Although this percentage might not seem especially high, it is a meaningful result in the case of a survey study of this nature in which many immeasurable factors are likely to shape subjects' attitudes.

Tables 2 and 3 report two different specifications of a regression model designed to estimate the effects of several causative factors on the ability of users to navigate their way around the catalog. The dependent variable is LOST, the user's rating of the catalog on a five-point scale ranging from 1 (felt lost) to 5 (felt I knew where I was). Again, because this variable has only five discrete values, an ordinal regression technique is used. Causative factors used in these regressions are similar to those included in table 1, although some less significant independent variables are now omitted. The only difference between table 2 and table 3 is that the latter table contains a smaller set of independent variables, omitting some of the least significant variables in table 2, namely EXPERIENCE, DANCE, and ESL.

In both tables 2 and 3, the SUCCESS coefficient is positive and statistically significant at the 1 percent level. Users who rated their catalog use as successful were significantly less likely to feel lost when searching the catalog than users who considered their search unsuccessful. In both tables, the AUTHOR and KEYWORD coefficients are positive and significant, at the 10 percent level in table 2 and at the 5 percent level in table 3. Users who implemented author or keyword searches were significantly less likely to feel lost relative to users who implemented subject searches (the latter being the control group).

In both tables the TIME coefficient is negative; it is marginally significant in table 2 and significant at the 10 percent level in table 3. The longer the time spent on the search, the more likely the user reported difficulty navigating the catalog. Additionally, in both tables the RESULTS coefficient is negative and significant at the 10 percent level, indicating that the greater the number of search results, the more likely the user felt lost.

The MUSIC and THEATER coefficients are negative in tables 2 and 3. In table 2 only the MUSIC coefficient is significant, however; and in table 3 only THEATER is significant. In both cases the significance level is 10 percent. These findings together suggest that users from the music and theater schools were more likely to encounter difficulty navigating the catalog relative to users from the art school (the control group), although these results are not especially robust as they are clearly sensitive to model specification.

Overall, the R^2 statistics indicate that the independent variables used in tables 2 and 3 explain 15 to 17 percent of the variation in perceptions across users, depending on the specification of the regression model. Again, although not especially high, these percentages are meaningful in the case of survey data.

Table 4 reports a regression model that estimates the effects of a number of causative factors on the user's perception of success.[32] The dependent variable in this table is SUCCESS, a dummy variable taking on a value of one if the user indicated that he or she had been successful in finding what he or she was looking for, zero otherwise. The independent variables are similar to those included in table 1, with the exception that TITLE KEYWORD, the least significant search type in table 1, is removed and replaced with SUBJECT, a dummy variable taking on a value of one if the user implemented a subject search,

zero otherwise.[33] A new variable, KNOWN ITEM, is introduced, which is a dummy variable taking on a value of one if the user implemented a known item search, zero otherwise. The only other difference between tables 1 and 4 is that the STAFF variable is not used in table 4.[34] Because the dependent variable in this regression is dichotomous, taking on a value of either one or zero, a logit estimation routine is used.[35]

Interestingly, none of the independent variables in table 4 are statistically significant at any conventional significance level, with the exception of theater school affiliation, which has a positive effect on success, significant at the 10 percent level. This finding means that search success or lack thereof is largely exogenous; i.e., not explained by any other factors included in the model. A user's perception of success or failure in the use of the catalog is independent of the other factors measured in the survey. This could mean that success is a state of mind, self-identified by the user and determined only by whether the user's session on the catalog accorded with his or her prior expectations.

Conclusions and Discussion

Our results indicate that users implementing subject searches are more likely to be dissatisfied with the organization of the catalog than users of known item searches. Moreover, users undertaking subject searches are more likely to report difficulty navigating the catalog than users of other types of searches. These findings are consistent with earlier catalog use studies, which consistently document problems with subject searching. Subject retrieval is clearly a difficult issue with which the library profession as a whole has been struggling for many years, and our results confirm that subject searching remains problematic in a Web environment.

Our study demonstrates that the most significant factor determining a user's attitude toward the Web catalog is the perception of whether or not the search was successful in finding what he or she was looking for. The success or failure of the search is the most important determinant of the user's judgment of catalog organization and of whether he or she felt comfortable using it. Users are more likely to assign high scores to catalog organization and their ability to find their way around the catalog if they perceive their search to have been successful than if they believe the search failed. However, the perception of success or failure itself is not explained by the other variables included in the study. Neither the user's level of experience with the catalog, the type of search conducted, nor the number of results retrieved explains the success of the search. This finding makes it difficult to identify specific improvements in the design of the Web catalog that would enhance the probability of user success. As indicated above, a user's perception of success appears to be largely subjective, driven primarily by the expectations that the user brings to his or her session on the catalog.

Therefore it will be critical to continue to emphasize catalog searching techniques in bibliographic instruction and training. Improvement in searching skills will ensure that users have realistic expectations of what they can achieve on the catalog and the ability to achieve these tasks in an optimal way. Much of the bibliographic instruction at CalArts is course-integrated. The ideal method of instruction is point-of-use—for the user to learn how to implement a search technique when he or she actually needs to use it. Such instruction is likely to be most effective in course-specific settings. For example, music students would be taught how to search for scores and sound recordings, while theater students would be taught how to search for plays in collections. Continued efforts in outcomes assessment will also be useful. Students and faculty who attend bibliographic instruction sessions could be asked to evaluate the extent to which their needs are being met.

The necessity of instruction is further reinforced by our finding that the longer the time spent on a search and the greater the number of search results, the more likely the user is to feel lost on the catalog. This result again suggests that instruction is needed to explain to users the most efficient way to conduct searches for different types of information.

Although our observation study did not explicitly identify system design changes, it did confirm for the library that the majority of users of the Web catalog were successful in finding what they were looking for, rated the organization of the catalog effective for finding information, and felt comfortable navigating the catalog.[36]

References and Notes

1. David W. Lewis, "Research on the Use of Online Catalogs and Its Implications for Library Practice," *Journal of Academic Librarianship* 13, no. 3 (July 1987): 152–57.
2. Ibid, 156.
3. Karen M. Drabenstott and Marjorie S. Weller, "Failure Analysis of Subject Searches in a Test of a New Design for Subject Access to Online Catalogs," *Journal of the American Society for Information Science* 47, no. 7 (July 1996): 519–37.
4. Ray R. Larson, "The Decline of Subject Searching: Long-Term Trends and Patterns of Index Use in an Online Catalog," *Journal of the American Society for Information Science* 42, no. 3 (Apr. 1991): 197–215.
5. Joseph R. Matthews, Gary S. Lawrence, and D. K. Ferguson, *Using Online Catalogs: A Nationwide Survey—A Report of a Study Sponsored by the Council on Library Resources* (New York: Neal-Schuman, 1983).

6. Carol A. Hert and Michael S. Nilan, "User-Based Information Retrieval System Interface Evaluation: An Examination of an On-Line Public Access Catalog," in *Proceedings of the 54th Annual Meeting of the American Society for Information Science* (Washington, D.C.: American Society for Information Science, 1991), 170–77.

7. Larson, "The Decline of Subject Searching," 208–10.

8. Nicholas J. Belkin et al., "Taking Account of User Tasks, Goals, and Behavior for the Design of Online Public Access Catalogs," in *Proceedings of the 53d Annual Meeting of the American Society for Information Science* (Toronto: American Society for Information Science, 1990), 69–79.

9. Sharon Seymour, "Online Public Access Catalog User Studies: A Review of Research Methodologies, March 1986–November 1989," *Library and Information Science Research* 13 (Apr.–June 1991): 89–102; Jon R. Hufford, "Use Studies and OPACS," *Technical Services Quarterly* 9, no. 1 (1991): 57–70.

10. Seymour, "Online Public Access Catalog User Studies," 100.

11. Pat Ensor, "User Characteristics of Keyword Searching in an OPAC," *College and Research Libraries* 53 (Jan. 1992): 72–80.

12. Ingrid Hsieh-Yee, "Student Use of Online Catalogs and Other Information Channels," *College and Research Libraries* 57 (Mar. 1996): 161–75.

13. Iljong Kenneth Park, "Comparing Major U.S. OPAC Systems for Developing Countries," *Libri* 47, no. 4 (Dec. 1997): 234–42.

14. Elisabeth Sinnott, "Fewer Errors Resulting from the Users' Misconception of the OPAC in 1992 than a Decade Ago: A Comparative Study of No Direct Hits and Zero Hits in Author Searches," *Cataloging and Classification Quarterly* 18, no. 1 (1993): 75–101.

15. Deborah D. Blecic et al., "A Longitudinal Study of the Effects of OPAC Screen Changes on Searching Behavior and Searcher Success," *College and Research Libraries* 60, no. 6 (Nov. 1999): 515–30.

16. Larson, "The Decline of Subject Searching."

17. We use the term "known item search" in the sense in which it is generally used in the literature, i.e., to refer to any search other than a subject search. A legitimate argument can be made, however, that keyword and even title word searches are frequently forms of subject searching and that an author search is not a known item search in the case of a search for additional works by a given author.

18. During summer and holiday breaks, most catalog users are library staff, faculty, and graduate students.

19. Stephen E. Wiberley Jr., Robert Allen Daugherty, and James A. Danowski, "User Persistence in Scanning Postings of a Computer-Driven Information System: LCS," *Library and Information Science Research* 12 (Oct.–Dec. 1990): 341–53.

20. Charles R. Hildreth, "The Use and Understanding of Keyword Searching in a University Online Catalog," *Information Technology and Libraries* 16 (June 1997): 52–62.

21. Alexandra Dimitroff, "Mental Models Theory and Search Outcome in a Bibliographic Retrieval System," *Library and Information Science Research* 14 (Apr.–June 1992): 141–56.

22. Drabenstott and Weller, "Failure Analysis of Subject Searches in a Test of a New Design for Subject Access to Online Catalogs."

23. Belkin et al., "Taking Account of User Tasks, Goals, and Behavior for the Design of Online Public Access Catalogs."

24. Hert and Nilan, "User-based Information Retrieval System Interface Evaluation."

25. Joan M. Cherry, "Improving Subject Access in OPACs: An Exploratory Study of Conversion of Users' Queries," *Journal of Academic Librarianship* 18, no. 2 (May 1992): 95–99.

26. Rosemary Thorne and Jo Bell Whitlach, "Patron Online Catalog Success," *College and Research Libraries* 55 (Nov. 1994): 479–97.

27. Judy Anderson, "Have Users Changed Their Style? A Survey of CD-ROM vs. OPAC Product Usage," *RQ* 34, no. 3 (spring 1995): 362–68.

28. These questions are similar to those in the exit questionnaire used by Belkin et al., "Taking Account of User Tasks, Goals, and Behavior for the Design of Online Public Access Catalogs."

29. On the calculation of required sample size in the case of a finite population, see Robert B. Jackson, Confidence Intervals and Sample Sizes for One Variable, 1999. Accessed Jan. 4, 2001, http://soaisweb.byu.edu/courses/ISYS580/Modules/stats/STA-CH10.HTM

30. Changing the default categories does not affect the remaining results.

31. The ordinal regression method takes account of the fact that respondents might not consider the differences between numerical answer choices to be equal. For example, a respondent might feel that the difference between a 4 and a 5 on a five-point scale is greater than the difference between a 2 and a 3. In an ordinal regression, what matters is the ordinal ranking of the responses, not the numerical distances between numbers. See Peter McCullagh, "Regression Models for Ordinal Data," *Journal of the Royal Statistical Society* 42, ser. B, no. 2 (1980): 109–42. Note that ordinal regression equations, such as those estimated in tables 1–3, do not contain a constant term. Table 4, which displays the results of a logit regression, does contain a constant term.

32. The regression in table 4 is just one example drawn from a series of regressions, varying slightly in specification, that all gave similar results.

33. The KEYWORD variable now becomes the control category for search type.

34. Staff make up only 1 percent of the sample.

35. The binomial logit model is based on the cumulative logistic function in which the natural logarithm of $p/(1 - p)$ is regressed on a linear function of the independent variables, where p is the probability that the binary dependent variable is equal to one. See Peter Kennedy, *A Guide to Econometrics*, 2d ed. (Cambridge, Mass.: MIT Pr., 1985), 189–202; Jan Kmenta, *Elements of Econometrics*, 2d ed. (New York: Macmillan, 1986), 550–53; and A. H. Studenmund and Henry J. Cassidy, *Using Econometrics: A Practical Guide* (Boston: Little, Brown, 1987), 174–75.

36. At the same time that the CalArts Library conducted this observation study, it also conducted a usability test of the Web catalog. The usability test, which required only eight participants, was more helpful in identifying screen design and navigational changes that could be made to clarify functions of the Web catalog and improve search success. The design changes identified through the usability study were put into effect in the summer of 2000 and a second round of usability testing is currently underway to determine whether these changes produced significant improvements in users' success.

Appendix A. CALIS User Study Observation Form

The Library is conducting a study about how the online catalog is used. We're looking at ways the catalog can be improved. I'd like to ask you to participate in the study because you're about to begin searching the catalog. Have you participated in this study before? (IF YES, thank the patron and find another subject, IF NO, continue).

I will observe you and take notes as you use the catalog. This is not a test of your skills, it is a test of how well our catalog works.

This should only take as long as you use the catalog, plus up to five minutes for you to complete a brief questionnaire about your experience.

Date: _____
Time started: _____
Time stopped: _____

Search #1

1. Search terms: _____
   _____
2. Area selected:
   Easy
   Advanced
   Databases
   Reserve Room
   My Account
3. Type of search:
   Author
   Author Keyword
   Instructor/course
   Title
   Title Keyword
   Databases: _____
   Subject
   Subject Keyword
   Keyword
   Keyword
   Limits/Numeric: _____
4. Number of results: _____
   Number of titles: _____

Comments: _____
_____
_____

Search #2

1. Search terms: _____
   _____
2. Area selected:
   Easy
   Advanced
   Databases
   Reserve Room
   My Account
3. Type of search:
   Author
   Author Keyword
   Instructor/course
   Title
   Title Keyword
   Databases: _____
   Subject
   Subject Keyword
   Keyword
   Keyword
   Limits/Numeric: _____
4. Number of results: _____
   Number of titles: _____

Comments: _____
_____
_____

Search #3

1. Search terms: _____
   _____
2. Area selected:
   Easy
   Advanced
   Databases
   Reserve Room
   My Account
3. Type of search:
   Author
   Author Keyword
   Instructor/course
   Title
   Title Keyword
   Databases: _____
   Subject
   Subject Keyword
   Keyword
   Keyword
   Limits/Numeric: _____
4. Number of results: _____
   Number of titles: _____

Comments: _____
_____
_____

Search #4

1. Search terms: _____
   _____
2. Area selected:
   Easy
   Advanced
   Databases
   Reserve Room
   My Account
3. Type of search:
   Author
   Author Keyword
   Instructor/course
   Title
   Title Keyword
   Databases: _____
   Subject
   Subject Keyword
   Keyword
   Keyword
   Limits/Numeric: _____
4. Number of results: _____
   Number of titles: _____

Comments: _____
_____
_____

Appendix B. CALIS User Study Questionnaire

1. How many years have you been at CalArts?
   - ☐ 1
   - ☐ 2
   - ☐ 3
   - ☐ 4
   - ☐ 5
   - ☐ 6
   - ☐ 7
   - ☐ 8–10
   - ☐ More than 10
2. Experience using catalog:
   - ☐ Constantly (5 or more times per week)
   - ☐ Frequently (2 to 4 times per week)
   - ☐ Occasionally (2 times per month)
   - ☐ Seldom (once a week)
   - ☐ Never
3. Are you:
   - ☐ BFA student
     Year: _____
     School/Program: _____
   - ☐ MFA student
     Year: _____
     School/Program: _____
   - ☐ Faculty
     School/Program: _____
   - ☐ Staff
     Department: _____
   - ☐ Alumni
   - ☐ Community/non-CalArts patron
4. Is English your native language?
   - ☐ Yes
   - ☐ No
5. When you searched the catalog today, please explain what you were looking for? Be specific: ___
   _____
   _____
6. Did you find what you were looking for?
   - ☐ Yes
   - ☐ No

   Comments on your catalog search: _____
   _____
   _____
7. The organization of the CalArts Library catalog is:
   - ☐ 1 Not effective for finding information
   - ☐ 2
   - ☐ 3
   - ☐ 4
   - ☐ 5 Very effective
8. When I was looking for information, I:
   - ☐ 1 Felt lost
   - ☐ 2
   - ☐ 3
   - ☐ 4
   - ☐ 5 Felt like I knew where I was

Additional comments: _____
_____
_____
_____

Analysis of Web-based Information Architecture in a University Library: Navigating for Known Items

David Robins and Sigrid Kelsey

This paper presents a descriptive study of the Louisiana State University Libraries' Web site. The intent of the study was to gain some idea of user demographics and satisfaction with the site at a given point in time and to test the site's navigation system. We wished to find out who was using the site, why they were using it, and to what extent they were satisfied with the site's navigation. We then assigned tasks (searching for known items) to subjects to better determine the extent to which the site's navigation system facilitated locating information on the site. Evaluation of the navigation system was based on a ratio of correct clicks to the sum of incorrect and back button clicks. This ratio may be compared to some predetermined optimal number of clicks needed to retrieve a known item. The implications of this research are both theoretical and practical. These models of in-house, Web-based information seeking may be used by other institutions of a similar nature that seek to provide useful Web sites for their users as well as to provide a basis for further research on the problem of Web-based development of information retrieval systems.

In fall 1999, the Louisiana State University (LSU) Libraries Web site contained more than six thousand layers: files ending with the extensions html, htm, or txt. The site provides access to databases, full-text journal articles and books, information about the library staff and hours, forms to reserve library classrooms, and more. With usage during the low time of day in 1999 greater than that for the peak time in 1995, it is obvious that the Web site is used more and more frequently for many purposes. In order to facilitate successful research through the libraries' Web site, it is necessary to analyze researchers' behavior when searching the Web site.

The LSU Libraries Web site was redesigned at the start of the spring 1999 semester. Prior to the redesign, the LSU Libraries Webgroup conducted library-wide forums open to all library employees. The Webgroup asked the forum participants to fill out a brief survey, discussing what they thought the major problems of the Web site were. Furthermore, the Webgroup presented various library Web sites to the participants, facilitating a discussion on the methods and styles of other library Web sites, and taking notes on the comments. Finally, the Webgroup presented the participants with several prototypes of a new Web site (first layer only) and solicited feedback.

Based on the feedback from the forums, five new prototypes for Web pages were designed, and the Webgroup asked the library employees for more feedback. The Webgroup research resulted in a new design for the LSU Libraries Web page, using a directory structure with seven headings: LOLA—online catalog, Electronic Resources, Key Links, General Information, Services, Library Department and Campus Libraries, and Search the Libraries' Web Pages.

The major problem specified at the forums was navigation: people did not always know where to click to find what they were looking for. The new design attempted to solve this problem in several ways: the new directory structure with related links under each heading made navigation more intuitive; a Search this Site feature, formerly available deeper within the site, was moved forward to the front layer; and an A to Z: Web Site Contents was added under the General Information heading to alphabetically list Web pages and topics likely to be searched for by Web site users.

Shortly after the new design was implemented, a link to the Web survey was added to the home page to survey users about the Web site design, initiating our research.

Problem

The ideal library Web site leads its patrons to whatever information they are seeking in a straightforward and efficient manner. The organization, wording, and content of the Web site are important in leading its users to desired information. Web pages can be organized in a number of ways: a directory structure is one of the most common ways to organize a site. Site maps, tables of contents, and search engines can help organize a Web site and facilitate easy navigation. The Web site developer must decide which of these materials to use, and how to best use them. In addition to the structure of the Web site, the developer must take into consideration the vocabulary on the site, such as library jargon. The following questions must be considered: What is library jargon? Are such terms as "indexes" and "databases" jargon? Some libraries have switched to using phrases such as "look for journal articles here." The library Web developer must decide whether eliminating such terms as "indexes" dumbs down the language of the libraries' site or enhances usability. This paper outlines some protocol library Web site designers can follow to ensure their Web

David Robins (drobins@pitt.edu) is Assistant Professor at the University of Pittsburgh School of Information Sciences; **Sigrid Kelsey** (skelsey@lsu.edu) is Electronic Reference Services and Web Development Coordinator at Louisiana State University Libraries, Baton Rouge.

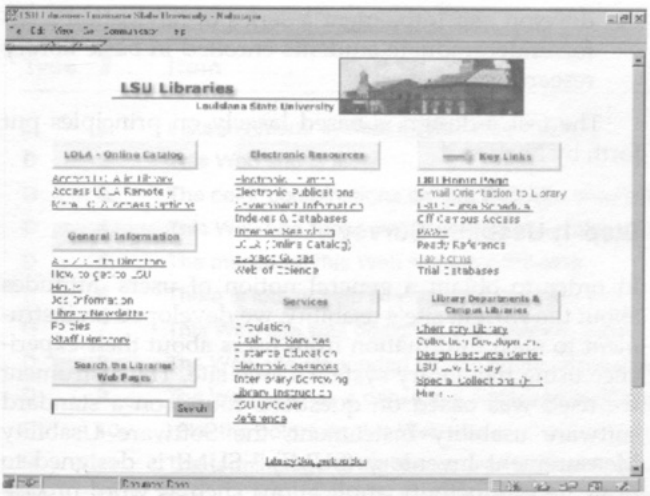


Figure 1. A screen shot of LSU Libraries Home Page, www.lib.lsu.edu. (Note: the site may have changed by the time this article is published.)

sites are providing contained information in a manner easily accessible to the library patrons.

Related Literature

The practice of creating Web-based information resources for libraries, or any other organization, is a new one. The Web has only been in existence for about twelve years, and the graphical browsers making it accessible to a broad base of users have only been available for about seven years. In addition, it has only been within the last four or five years that Web site development tools have made it possible for developers without sophisticated technological training and experience to create Web sites. The art of Web site development is in its infancy.

User-behavior studies became prevalent in the early nineties, usually regarding CD-ROMs. Many early studies survey the amount of success the users have in searching various databases; some of these discuss redesigning interfaces based on the user behavior. Many of these studies are listed and abstracted in "User's Information-Seeking Behavior: What Are They Really Doing? A Bibliography."[1]

Some early studies include the one conducted by Puttapithakporn in 1990, which surveys twenty-three students in an information science class searching ERIC on CD-ROM.[2] The focus of the article presents problems the students encountered. Bucknall and Mangrum study the CD-ROM service at the University of North Carolina at Chapel Hill, concluding that librarians need to be prepared to instruct first-time users as well as respond to complex search questions, and that most users prefer staff assistance to other forms of help.[3] Culbertson's study uses Total Recall, software created by Computer Foundations, to capture the keystrokes of CD-ROM users.[4] Culbertson notes that users rarely use advanced searching options, and recommends user training.

In the mid-nineties, user studies regarding systems created within the library begin to appear in the literature. The conclusions mention redesigning interfaces rather than focusing on better user training. Catledge and Pitkow present an early study of user behavior with recommendations for Web interfaces in their paper "Characterizing Browsing Strategies in the World Wide Web."[5] The authors use a log analysis to recommend interface features, such as placing must-see information within two to three jumps of the initial home page, and using indexes throughout the site.

In a 1998 article, Carter examines the Indiana University–Purdue University, Indianapolis (IUPUI) library Web site interface, initially surveying staff about the site, and following up with a task questionnaire similar to the one we sent out to students.[6] Carter asked library staff familiar with the interface to complete the questionnaire; we asked LIS 1001 students, many of whom had never used the Web site, to complete ours. Carter counted the number of clicks for each task, also similar to our study. Finally, Carter sent out a survey to the library staff comparable to the survey we posted on our Web site. She concludes with some basic tips, such as having a link back to the main menu, and creating a broad menu structure.

Veldof, Prasse, and Mills discuss two types of usability evaluation methods, and recommend a combination of the two: studies with real users, and ones without real users, such as applying a set of heuristics to the design.[7] In our studies, we have combined several types of both methods. Jakob Nielsen's usability engineering principles and the three-click rule are examples of heuristics.[8]

While some principles of design are available for librarians to follow as they attempt to provide Web-based information resources for their users, and there are some empirical studies that provide a solid basis for design, the area of research is a new one. Furthermore, academic libraries provide various levels of complexity that a Web site must address. For example, these sites must provide links to bibliographic databases; the library's Online Public Access Catalog (OPAC); general information about the library, such as hours, services, and departments; and internally created resources, such as Web pathfinders for specific subject areas.

General Web design guidelines, or heuristics, offer useful tips and information to begin evaluating a Web site. Nielsen's Top Ten Mistakes in Web Design, for example,

provides some clear, simple guidelines every Web site should strive to implement, such as avoiding complex URLs and nonstandard link colors.[9]

Rosenfeld and Morville identify a wide variety of issues related to what can be called information architecture (IA).[10] IA is a term that has come to represent efforts to develop best practices for the design of Web-based information sources. Navigation systems, taxonomies, and other organization systems, search systems, labeling systems, and overall site design are all subsets of IA. Our focus here is on navigation, but as Rosenfeld and Morville point out, navigation systems are inextricably tied to such subsystems as labeling and organization. Robins presents another explanation of IA as a tool for records management.[11]

Navigation in large Web sites presents challenges for Web site designers. Some of the issues related to navigation include:

- labeling: for example, whether one should use precise, technical language, or labels more appropriate for the nontechnical user;
- hierarchy and context of navigation systems: which pieces of the navigation system should appear on every page, and which pieces should appear only on some pages; and
- breadth and depth of menu systems.

Dietrich, Gordon, and Wexler examine the issue of breadth and depth of Web-based menus and find that users make fewer errors when menu systems are broad rather than deep.[12] Similarly, Morkes and Nielsen find that most Web site users (79 percent in their study) do not read content word by word, but rather scan pages.[13] Presumably, the users are looking for something specific that they suspect might be on the page or something less well-defined (i.e., they may be operating under the assumption that they will know what they are looking for when they see it). In any case, a broad menu system allows a user to easily scan its contents for the path to desired information by allowing a user to rely on recognition of the desired link as opposed to attempting to recall or decipher a more general heading.

Based on our findings of similar studies and related writings, we decided to test the effectiveness of LSU Libraries' navigation system design.

Research Design

In order to study the effectiveness of our Web site's navigation system, we chose to:

- conduct a Web survey to determine the current attitudes toward site usability; and
- develop and implement a series of navigation tasks for undergraduate students enrolled in basic library research courses.

The task's design is based largely on principles put forth by Nielsen.[14]

Step 1: Usability Survey

In order to obtain a general notion of users' attitudes about the current site's usability, we developed an instrument to elicit information from users about their experience using the library system's Web site. The instrument we used was based on questions found on a standard software usability instrument, the Software Usability Measurement Inventory (SUMI).[15] SUMI is designed to evaluate productivity applications such as word-processors and spreadsheets.

Because we were studying the usability of a Web site, with which users interact somewhat differently than they do with productivity software, we made certain modifications to the instrument. For example, SUMI asks subjects to respond "agree," "undecided," or "disagree" to the following statement: "It takes too long to learn the software commands." In this case we decided not to use the statement in the survey because there are no commands to learn on the Web site. Similarly, we substituted "Web site" for "software" to make the statements more appropriate for our purposes. Finally, we chose not to use SUMI's three-tiered response system. Using the instrument, the respondents rated their level of agreement on a Likert scale of 1 to 5 (1 = strongly agree, 5 = strongly disagree). We chose to have subjects scale their responses so that we could get a more general picture of their attitudes toward the site's usability. We hoped to gain from the administration of the instrument a better understanding of navigation and discovery, design, and overall feelings.

Items in the navigation and discovery section sought to elicit attitudes about the respondents' ability to find what they were looking for. Furthermore, we wanted to find out if users found unexpected, serendipitous information resources by using the site. Finally, some of the items in this section sought to find out if the terminology used on the site was a problem. The second category of survey items dealt with the site's design: how fast the site loaded; the consistency of layout, graphics, and headings; and colors. We wanted to know if the Web site looked and felt like other Web sites with which people were familiar. The third category attempted to obtain users' overall attitudes and feelings about using the Web site. Table 1 shows the survey instrument with items listed according to purpose.

Table 1. Questionnaire Sorted by Question Type

Type	#	Item
D	1.	I like the menu system on this Web site
D	2.	This Web site is slow
D	3.	The color combinations on this Web site should be changed
D	4.	This Web site "behaves" like most other Web sites
D	5.	The menus on this Web site are confusing
D	6.	There is lots of help available on this Web site
D	7.	This Web site was designed with users in mind
D	8.	The Web site responded rapidly during navigation
D	9.	The Web site has a consistent "look and feel" throughout
D	10.	I like the colors on this Web site
N	11.	The terminology in the menus was familiar to me
N	12.	I always knew where to find what I was looking for
N	13.	I found ONLY the information I was looking for and nothing else
N	14.	I had to go through too many menus to find what I was looking for
N	15.	The search mechanism on the Web site was helpful
N	16.	It takes too long to find something on this Web site
N	17.	I sometimes wonder whether I'm clicking on the right menu
N	18.	The Web site provides the opportunity to discover information
N	19.	The menus on this Web site are logically constructed
N	20.	Navigation of this site was problematic
N	21.	The categories on the main page provided a guide to the information I needed
N	22.	I unexpectedly found useful information on this Web site
N	23.	It was easy to find specific information on this Web site
O	24.	I needed help finding what I needed on this Web site
O	25.	Using this Web site was fun
O	26.	I felt frustrated using the Web site
O	27.	It took too long to find what I wanted on this site
O	28.	I felt lost on this Web site
O	29.	Going to the library makes me uncomfortable
O	30.	I enjoyed using this Web site
O	31.	I would recommend this Web site to my colleagues
O	32.	I felt tense using this Web site
O	33.	There was too much jargon in the menus
O	34.	Using the library, in general, can be frustrating
O	35.	I'll definitely use this site in the future

D = Design, N = Navigation/discovery, O = Overall feeling

Step 2: Navigation/Usability Tasks

In order to observe how people use the Web site, and to determine navigational problems associated with the site, we set up a number of tasks for users to perform. The tasks were all searches for information known to exist on the Web site. For this study, 314 undergraduate students in various sections of LIS 1001 (Library Research Methods and Materials) were asked to perform searches. Each student was assigned two tasks, and students who completed the assignment received extra credit in the course. Because the assignments were distributed during the first week of class (before the students were taught about the Web site in class), the students completing the tasks represented a cross section of undergraduates from a variety of departments. The variance in the distribution of each task is due to how the assignments were handed out—in some classes, they were not shuffled before handing out, so that more students received the same assignments. Table 2 shows the breakdown of tasks, the number of students involved, and the minimum number of navigational moves necessary to complete each task. Each subject was asked to document each navigational move they made during their attempt to complete their assigned tasks. A navigational move was defined as a click on a hyperlink.

Analysis of these tasks proceeded on two conventions. First, the starting point of each task was the LSU Libraries Web Site. Therefore, any navigation necessary to get to the LSU Libraries Web site was not counted as a move necessary to complete a task. For example, if a subject documented that they went to the main LSU Web site first and then used three clicks to get to the LSU Libraries Web site, and then proceeded to complete the assigned task, the first three navigational moves were not counted. The second convention used to analyze navigation data was that finding the link to the desired information constituted completion of the task. That is, clicking on the link

Table 2. Navigation Tasks Assigned to Subjects

Task number	Task	Number of potential subjects	Minimum number of moves to complete task
1	How does one reserve one of the library's electronic classrooms?	20	2
2	Who is the Dean of Special Collections?	21	2
3	What is Sigrid Kelsey's e-mail address?	36	1
4	Find what Y2K resources are at LSU Library	20	1
5	Find a page containing databases related to art	20	1
6	Find a link to the Medline database	37	1
7	Find a link to Project Muse	43	2
8	Find the phone number to call to renew a book	43	1
9	Find the call number to *Vanity Fair* by William Makepeace Thackeray	37	1
10	What time does the Design Library close on weekdays?	37	2
	Total	314	

Note: The difference in the number of subjects assigned to each task is due to the fact that subjects were drawn from several sections of a class (LIS 1001 (Library Research Methods and Materials)). Each section varied considerably in number of students enrolled. Subjects are "potential" subjects because participation was not compulsory. Rather, extra credit in the course was the only inducement offered for participation.

to the page containing the desired information was not counted in the total navigational moves necessary to complete a task.

Another consideration in analyzing navigational moves was the minimum number of moves necessary to complete a task. In a Web site of any complexity, there may not be one right way to retrieve a known information item. In fact, on most of the assigned tasks, we identified more than one way to navigate to the desired information. Therefore, it is more appropriate to talk about the optimum or minimum number of moves. In some cases, it was possible to navigate various paths from the libraries' home page to the desired page and still complete the task in the minimum number of moves.

Finally, in analyzing subjects' navigation of the site, we used the coding scheme shown in table 3. Each move documented by subjects was coded according to the scheme presented in table 3, and totals for each were counted. This phase of the study was designed to address research questions 4 and 5 dealing with navigational processes and whether users were successful.

Results

The results of this study are presented in the order that they address the research questions. Results are also presented to reflect the stages specified in the research design. The usability survey in step one addressed the research questions "Who is using the LSU Libraries' Web site?", "For what purposes do users consult the LSU Libraries'

Table 3. Coding Scheme Used to Analyze Tasks Assigned to Subjects

Code Abbreviation	Code name	Description
I	Incorrect	Any navigational move that will not directly lead to the specified known item
B	Back one page, or Home	Clicking any "back" or "home" navigational device
C	Correct	Any navigational move leading directly toward the specified known item
O	Offsite move	Using a source outside of the LSU Libraries Web site to find the desired information

Web site?", and "What impressions about the site do its users have?" The task analysis in step two addressed the research questions In what ways do they navigate the site? How successful are users when attempting to find desired information on the LSU Libraries Web site?

Step 1: Survey Results

First, we will show what types of users use the LSU Libraries Web site, for what purposes they use it, and what impressions they have about it. In total, we had 129 responses to the survey, but many of the subjects had to be deleted because they did not respond to a majority of the items on the survey. We made the determination that if a respondent left only two or three items blank, we would tally his or her responses. Therefore, the totals in table 4 (indicated under "n") are not always the same, but the range among those totals is only three (mean = 64, median = 64, mode = 63, standard error = 0.4).

The respondents represent a self-selected group because we solicited responses in the form of a request posted on the Web site itself. In the next section, we describe the respondents' demographics.

Who is using the LSU Libraries' Web site, and how are they using it?

Before discussing the results of our survey, we must state the following caveats. First, the respondents were not selected at random. Rather, they simply responded to a request on the Web site to fill out the survey. As such, respondents were self-selected. Second, and owing to the fact that respondents were not under any sort of control or obligation to the researchers, there is no way to assure that respondents were being honest or accurate in their responses. Third, the respondents heavily represent the undergraduate population, so any deviation from the mean is, in essence, a deviation from the undergraduate mean. In any case, our goal with the survey was to get a general feeling about how the users of this Web site respond to using it. It is, therefore, in the spirit of gaining baseline data that we proceeded with the survey.

In a three-month period, seventy-four respondents filled out the survey. Out of the seventy-four respondents, thirty-two (43 percent) were undergraduate students, fifteen (20 percent) were not affiliated with LSU, fourteen (19 percent) were graduate students, six (8 percent) were faculty, and five (7 percent) were staff. Two (3 percent) did not report their status. There were not enough responses in each respondent class to determine if differences were real or biased. We did observe some tendencies, which should, of course, be taken as possible future areas of research and not as conclusive results. For example, faculty were more likely than students to disagree with the statement, "The Web site provides the opportunity to discover information." Faculty agreed at a higher rate than students that the site was jargon laden. However, since there were only six responses from faculty, these results are tentative at best.

Nevertheless, the purposes for which the site was used were somewhat varied (see the summary in Table 5). Of all respondents who indicated their primary purpose for visiting the site (55 in all), 24 were just browsing, 24 were working on a specific assignment, and 5 were working on a dissertation or thesis. Most of the undergraduates (28 indicated purpose) used the site for assignments (17) and the rest used it simply for browsing (11). The 14 graduate students who indicated use were almost evenly divided among just browsing (4), assignments (5), and dissertation/thesis (5). Those not affiliated with LSU and indicating use (6) were all browsing. Only one faculty member and two staff members reported browsing and one staff member reported working on an assignment.

What impressions about the site do its users have?

The results varied greatly, from comments such as this one, from an English professor:

> I consider the LSU Library Home Page "impossible." It is highly inefficient. There are too mahy [sic] fine distinctions of primarily administrative interest. Especially perplexing is the huge list of possible interfaces that one receives when logging on from outside the library or campus. It might be acceptable if some of these interfaces worked, but, alas, they are all either impossibly slow or crash the host system (or crash it in virtue of being slow and timing it out). I am an experienced computer user and have used electronic library catalogues all over the world and have never seen such a miasma. Please simplify the system and arrange it intuitively rather than on the basis of the arrangements you have with software and database vendors. I have been working with this system now for 45 minutes trying to get a single citation and have now given up in order to TELNET to other remote libraries using my own independent browser.

There were also more positive comments, such as the following, from a graduate student in library and information science:

> I feel that this is a fine attempt towards the ideal of a totally intutative [sic] and very user-friendly Web site. I am very, very pleased at the efforts laid out in this cyber-document. Please continue to maintain it.

Survey respondents rated their level of agreement with statements about the site's navigability and usability, its design, and their overall feelings about using it. Table 4 summarizes the results of these rankings.

The Web survey generated some basic idea about how people searched the Web site and what they did and didn't like. Because the nature of Web sites is dynamic, the Web site did change over the timespan in which the Web survey was offered. However, no drastic changes in

Table 4. Survey Results Presented in Raw Numbers and As a Percentage for Each Item

Type†	Item‡	Question	Rating*							Percentage*				
			1	2	3	4	5	n	Mean	1	2	3	4	5
D	2	I like the menu system on this Web site	14	18	16	9	12	69	2.81	20	26	23	13	17
D	3	This Web site is slow	16	12	15	20	6	69	2.83	23	17	22	29	9
D	4	The color combinations on this Web site should be changed	10	4	16	18	21	69	3.52	14	6	23	26	30
D	5	This Web site "behaves" like most other Web sites	11	19	24	5	9	68	2.74	16	28	35	7	13
D	12	The menus on this Web site are confusing	14	10	18	15	7	64	2.86	22	16	28	23	11
D	13	There is lots of help available on this Web site	8	7	28	11	10	64	3.13	13	11	44	17	16
D	20	This Web site was designed with users in mind	10	21	16	6	11	64	2.80	16	33	25	9	17
D	23	The Web site responded rapidly during navigation	9	20	19	9	7	64	2.72	14	31	30	14	11
D	26	The Web site has a consistent "look and feel" throughout	12	21	19	5	5	62	2.48	19	34	31	8	8
D	35	I like the colors on this Web site	14	16	15	7	9	61	2.69	23	26	25	11	15
N	1	The terminology in the menus was familiar to me	15	23	16	4	11	69	2.61	22	33	23	6	16
N	6	I always knew where to find what I was looking for	7	13	12	17	18	67	3.34	10	19	18	25	27
N	7	I found ONLY the information I was looking for and nothing else	4	6	25	9	22	66	3.59	6	9	38	14	33
N	9	I had to go through too many menus to find what I was looking for	20	16	12	9	10	67	2.60	30	24	18	13	15
N	11	The search mechanism on the Web site was helpful	10	18	18	7	11	64	2.86	16	28	28	11	17
N	16	It takes too long to find something on this Web site	15	13	12	14	9	63	2.78	24	21	19	22	14
N	17	I sometimes wonder whether I'm clicking on the right menu	20	12	13	10	9	64	2.63	31	19	20	16	14
N	19	The Web site provides the opportunity to discover information	18	19	14	3	9	63	2.46	29	30	22	5	14
N	22	The menus on this Web site are logically constructed	6	26	17	8	6	63	2.71	10	41	27	13	10
N	25	Navigation of this site was problematic	10	15	16	13	8	62	2.90	16	24	26	21	13
N	31	The categories on the main page provided a guide to the information I needed	10	22	11	10	9	62	2.77	16	35	18	16	15
N	32	I unexpectedly found useful information on this Web site	9	15	23	8	8	63	2.86	14	24	37	13	13
N	34	It was easy to find specific information on this Web site	7	18	18	10	9	62	2.94	11	29	29	16	15
O	8	I needed help finding what I needed on this Web site	14	15	25	12	0	66	2.53	21	23	38	18	0
O	10	Using this Web site was fun	8	7	23	8	20	66	3.38	12	11	35	12	30
O	14	I felt frustrated using the Web site	18	10	8	14	14	64	2.94	28	16	13	22	22
O	15	It took too long to find what I wanted on this site	18	11	12	12	10	63	2.76	29	17	19	19	16
O	18	I felt lost on this Web site	10	15	12	13	13	63	3.06	16	24	19	21	21
O	21	Going to the library makes me uncomfortable	9	6	14	5	29	63	3.62	14	10	22	8	46
O	24	I enjoyed using this Web site	8	17	19	5	14	63	3.00	13	27	30	8	22
O	27	I would recommend this Web site to my colleagues	13	16	18	8	8	63	2.71	21	25	29	13	13
O	28	I felt tense using this Web site	9	6	18	15	16	64	3.36	14	9	28	23	25
O	29	There was too much jargon in the menus	10	4	19	17	12	62	3.27	16	6	31	27	19
O	30	Using the library, in general, can be frustrating	18	9	15	13	8	63	2.75	29	14	24	21	13
O	33	I'll definitely use this site in the future	26	17	9	1	10	63	2.24	41	27	14	2	16

* Items were rated from 1 (strongly agree) to 5 (strongly disagree).

† D = design-related question, N = navigation/discovery question, and O = overall feelings.

‡ The numbers in the Item column are the order in which the items appeared on the survey as presented to respondents.

the look and feel or organization of the Web site were implemented during this time. While this survey gave us some general ideas about user attitudes toward our Web site, it did not provide any specific information unless the respondents commented in such a way. Because of the Likert scale, the survey recorded general feelings but not the reasons behind the answers. Moreover, many users filled out some answers when it was clear that their main objective was to request a book or complain about the facilities. Nevertheless, the information from the survey gave us direction in formulating questions for the second phase of the study; for example, 48 percent of the respondents agreed (chose 1 or 2) that it took too long to find what they were looking for on the Web site.

Step 2: Task Analysis

We wanted to see how the Web site's users navigated in order to find information they wanted. To do so, we constructed a set of tasks for users to accomplish. Users were instructed to find specific, known items in the site. In all, there were ten tasks given to various sets of users, with each user receiving two tasks. The variance in the number of users completing various tasks is due in part to uneven distribution. In some of the classes, the assignments were not properly shuffled before handing them out. Nevertheless, the percentages in the results give a clear idea of which tasks were completed with the greatest success. Because the menu system on this site is quite broad as opposed to deep, it was possible to assign tasks that could be accomplished in one or two moves. Therefore, it was easy to identify when items were found in an optimal way. There were times when subjects found alternate means of getting to the known items using the optimal number of moves, but these cases were rare. Ordinarily, a user needed to take one path to the known item in order to get there using the minimal number of moves. Moves are defined as the point at which a subject made use of a hyperlink to move from one point in the Web site to another.

Subjects recorded each move, correct or incorrect, and the researchers coded the moves using a simple scheme: C = correct move; I = incorrect move; B = clicking to go back a screen; and O = offsite browsing. An example of offsite browsing was when subjects used the LSU directory, or a 411 service, to find an e-mail address, as opposed to using the LSU Libraries' Web site to do so. They found the information, and while that is not strictly incorrect, we were interested in the libraries site's navigation system, not in the subject's ability to find information on the Web at large.

These categories were derived by analyzing samples of moves, and were not decided upon a priori by the researchers. This scheme allowed us to:

- compare patterns of moves among different tasks;
- compare patterns of moves within tasks; and
- compare optimal moves with observed moves.

In what ways do users navigate the site?

We began the task analysis by looking at various comparisons of moves within and among tasks. In all, we observed 772 moves by all subjects across all tasks. Total moves within each task and for each category within tasks are shown in table 6. In part, the data shown in table 6 address the question, "What percentage of move types were observed for each task?" Here, we see that the task to find a link to Project Muse appears to have been particularly troublesome. Forty-three subjects made 159 moves in completing the task, an average of 3.7 moves to accomplish what could have been done in 2 moves. Interestingly, the same subjects performed the second most efficient task (1.81 moves per subject) on the telephone renewal question. Therefore, finding the link to Project Muse was difficult. Perhaps the difficulty was in the fact that the students did not know what Project Muse was and, consequently, did not know where to begin looking.

In addition, table 6 shows that 65 percent of all moves were correct moves. This probably speaks well for the

Table 5. Purposes for Which Users Visited the LSU Libraries Site

Purpose	Undergrad	Grad	Prof.	Staff	Not Affiliated	Blank	Total
Assignments	17	5	--	1	--	1	24
Browsing	11	4	1	2	6	--	24
Dissertation	0	5	--	--	--	--	5
Research	0	--	1	--	--	--	1
Work	0	--	--	1	--	--	1
Total	28	14	2	4	6	1	55

Web site in general. However, the percentage of correct moves ranged from 91 percent (for the design library closure time task) to 44 percent (for the e-mail task). It is unclear what this large range suggests, but it is probably due, at least in part, to the fact that efficient navigation relies on language to guide users to needed information. Language is notoriously ambiguous, and users will have problems making navigation choices; therefore much thought should go into labeling design.[16]

One point of interest in the data shown in the following tables is that two of the tasks, Sigrid Kelsey's e-mail and Project Muse, instigated a high percentage of off-site searches by users. We will see that Project Muse was particularly troublesome for subjects in other ways, but the e-mail address task was not otherwise difficult. Subjects may have been familiar with other services from which e-mail addresses could be found, and immediately sought out those resources. An additional point of confusion may have been that the users were not aware that Sigrid Kelsey is a libraries' employee, even though she introduced herself while distributing the assignments.

Another question we might ask about how users navigate the site is, "Of all moves observed across all tasks, what percentage occurred in each category?" This analysis will enable us to see which tasks show unusually high occurrences in each category. Table 7 addresses this question. The Project Muse, art databases, and Medline tasks accounted for 53 percent of all incorrect moves, 23 percent, 17 percent, and 13 percent respectively. All other tasks accounted for percentages ranging from 2 percent to 11 percent. However, since more than twice as many subjects were assigned to the Project Muse task compared to the art databases task (forty-three and twenty respectively), the latter seemed to have led subjects astray more often.

The percentages of correct moves were relatively evenly distributed among the tasks, with the exception of the reserving electronic classroom, Sigrid Kelsey's e-mail, Y2K resources, and the art databases tasks. These four tasks accounted for only 22 percent of all correct moves, and in each of these tasks, correct moves were equal to or below the percentages of incorrect moves for those tasks. These percentages are another indicator that subjects may have had an unusual amount of difficulty with these tasks, and may point to areas in which the Web site may need redesign.

How successfully do users navigate the site?

Obviously, it is difficult to compare the ten tasks as equals. Two factors complicate the issue. One is that the tasks themselves are different. The other is that the same groups did not do all ten tasks—each participant had only two tasks. That being said, we can make some comparisons by calculating the average moves per task made by each subject. This analysis will show whether there are vast differences among tasks regarding each task's difficulty. It will also provide a measure of the efficiency with which subjects accomplished each task. In table 8, we show data that rank the efficiency with which subjects accomplished each task, measured as a ratio of correct

Table 6. Moves by Category and Task, with Percentages of Moves within Task Categories

		Move									
		Incorrect		Back/home		Correct		Off-site		Total	
Task	No. of subjects	No.	%	No.	%	No.	%	No.	%	No.	%
Reserving electronic classroom	14	11	23	7	15	29	62	0	0	47	100
Dean of Special Collections	21	11	16	9	13	50	71	0	0	70	100
Sigrid Kelsey's e-mail	36	6	10	5	8	27	44	23	38	61	100
Y2K resources at LSU	20	11	30	8	22	17	46	1	3	37	100
Find art databases	20	21	30	11	15	39	55	0	0	71	100
Find link to Medline	37	16	17	10	11	64	67	5	5	95	100
Link to Project Muse	43	28	17	17	11	81	50	33	20	159	100
Find number for phone renewal	43	10	13	3	4	65	83	0	0	78	100
Call number to *Vanity Fair* (book)	37	6	8	8	11	58	77	3	4	75	100
Time design library closes	37	2	3	1	1	72	91	4	5	79	100
Total	308	122	16	79	10	502	65	69	9	772	100

Table 7. Categories and Tasks Expressed As a Percentage of Moves within Categories across Tasks

Task	No. of subjects	Move: Incorrect No.	Incorrect %	Back/home No.	Back/home %	Correct No.	Correct %	Off-site No.	Off-site %	Total
Reserving electronic classroom	14	11	9	7	9	29	6	0	0	47
Dean of Special Collections	21	11	9	9	11	50	10	0	0	70
Sigrid Kelsey's e-mail	36	6	5	5	6	27	5	23	33	61
Y2k resources at LSU	20	11	9	8	10	17	3	1	1	37
Find art databases	20	21	17	11	14	39	8	0	0	71
Find link to Medline	37	16	13	10	13	64	13	5	7	95
Link to Project Muse	43	28	23	17	22	81	16	33	48	159
Find number for phone renewal	43	10	8	3	4	65	13	0	0	78
Call number to *Vanity Fair* (book)	37	6	5	8	10	58	12	3	4	75
Time design library closes	37	2	2	1	1	72	14	4	6	79
Total	308	122	100	79	100	502	100	69	100	772

moves to the sum of incorrect and back moves (C/(I+B)) in each task. That is, we combined back moves with incorrect moves under the assumption that back moves were the result of some mistake, real or perceived, that inclined the subject to start over.

Of note is the fact that the design library closure question was by far the most efficient search by our measure, a 24 to 1 ratio of correct to incorrect and back moves. The average ratio across all tasks was 2.51. The rest of the tasks' ratios were much closer to the average, and seven of the ten tasks were at or below the across task average. Being below average, by our measure, means that the search was less efficient, demonstrating a tendency for subjects to make fewer correct moves or for them to make more incorrect and back moves. Finding art databases and Y2K resources were the least efficient, according to this measure. Once users went off the libraries' site, the circumstances changed, so we did not include off-site moves in this formula.

Another way to examine the efficiency with which subjects navigated the site is to ask, "How many moves, on average, did each subject make during a task?" That is, we address the issue of how closely the subjects were able to approach the optimum number of moves for each task. In addition, this question normalizes the effects of varying numbers of subjects in each task. As in table 8, we used table 9 to rank the tasks from most efficient to least, and similar results occurred, but not exactly the same. It should be remembered that this measure was based purely on how many moves were made per subject, and not what kind. Therefore, efficiency by this measure is closely linked to the optimal number of moves for each task. The measure on which table 8 is sorted is moves-per-subject minus optimal number of moves. Table 8 shows that the fewest number of moves-per-subject was done finding Sigrid Kelsey's e-mail address, and the most were spent on finding the link to Project Muse. None of the tasks were accomplished, on average, in the optimal number of moves, although individual subjects did routinely do tasks using optimal strategy.

It is illuminating to compare tables 8 and 9 as efficiency measures. For example, the e-mail question was about an average C/(I+B) ratio, and yet, it was the most efficient in terms of average moves per subject. At the same time, the design library closure time task was highest on the C/(I+B) ratio and dropped to the middle of moves per subject ranking. This is due to the fact that these are two different measures of efficiency. One is measuring correct versus incorrect moves in each task, and the other is measuring the average moves that subject made during each task. It is recommended that analysts should refer back to the raw numbers (see tables 6 and 7, both of which express raw counts) to ascertain why differences occur. In the case of the e-mail and design library questions, it should be remembered that subjects only made a total of two incorrect moves and one back move, while making seventy-two correct moves. The e-mail question's subjects made six incorrect, five back, and twenty-seven correct moves. This indicates that, when considering redesign questions, the part of the navigation system that leads to e-mail addresses should be examined more closely than that which leads to hours of operation for branches.

Table 8. Ratio of Correct (C) to Incorrect (I) Moves and Back (B) Moves

Task	N	Correct/incorrect or back	Avg. correct/incorrect or back	Difference
Time design library closes	37	24.00	2.51	21.49
Find number for phone renewal	43	5.00	2.51	2.49
Call number to *Vanity Fair* (book)	37	4.14	2.51	1.63
Dean of Special Collections	21	2.50	2.51	-0.01
Find link to Medline	37	2.46	2.51	-0.05
Sigrid Kelsey's e-mail	36	2.45	2.51	-0.06
Link to Project Muse	43	1.80	2.51	-0.71
Reserving electronic classroom	14	1.61	2.51	-0.90
Find art databases	20	1.22	2.51	-1.29
Y2K resources at LSU	20	0.89	2.51	-1.62
Total	308			

Table 9. Comparison of Within-Task Differences between Moves per Subject and Mean Moves per Subject

	N	Total	Column A: Avg. moves/ subj. within task	Column B: Optimal no. of moves/task	Column A minus Column B	Column C: Avg. moves/ subj. across tasks	Column A minus Column C
Sigrid Kelsey's e-mail	36	61	1.69	2	-0.31	2.51	-0.81
Call number to *Vanity Fair* (book)	37	75	2.03	2	0.03	2.51	-0.48
Find number for phone renewal	43	78	1.81	1	0.81	2.51	-0.69
Y2K resources at LSU	20	37	1.85	1	0.85	2.51	-0.66
Time design library closes	37	79	2.14	1	1.14	2.51	-0.37
Find art databases	20	71	3.55	2	1.55	2.51	1.04
Find link to Medline	37	95	2.57	1	1.57	2.51	0.06
Link to Project Muse	43	159	3.70	2	1.70	2.51	1.19
Dean of Special Collections	21	70	3.33	1	2.33	2.51	0.83
Reserving electronic classroom	14	47	3.36	1	2.36	2.51	0.85
Total	308	772	2.51				

Discussion and Conclusions

This study was conducted to find out more about how well a particular Web site's navigation system works for various people doing various tasks. We conducted a two-phased study in which: (1) users were surveyed regarding their opinions and feelings about the Web site in general and its navigation system in particular; and (2) users were assigned tasks to test the Web site's navigability. Two measures of navigation efficiency were developed in the course of the study: (1) correct to non-correct moves for tasks; and (2) average moves to optimum moves comparison. We found that some parts of the site's navigation system

worked better than others. It should be noted that this study indicates what parts of a site's navigation system should be investigated and what parts might be left alone. This is an import-ant question when studying a site's information architecture.

The two phases of our study helped identify problems, and may be used in the analysis of other sites. The first phase identified areas in which there may be problems. For example, the quote that stated dissatisfaction with the site provided us with specific information about sources of problems in the site. In addition, while the survey items suggested that users were generally satisfied with the site, there were also indications that some parts of the site needed attention. For example, consider the following items and average results (on a 1 to 5 scale, 1 = strongly agree, 5 = strongly disagree)

- (item 6) I always knew where to find what I was looking for (4.2)
- (item 10) Using this Web site was fun (4.3)
- (item 13) There is lots of help available on this Web site (4.0)
- (item 34) It was easy to find specific information on this Web site (3.6).

Items 6, 10, and 34 are directly related to site navigation. These items do not give overwhelming support to the notion that the site is easy to navigate. Item 10 is mentioned only as another way to subjectively evaluate the overall effectiveness of the site.

Based on the results of the survey and assignment, several changes were made to the Web site. Approximately three hundred out-of-date files were removed so that they no longer show up in the results of a Search this Site search. A dynamic database listing all the databases, indexes, e-journals, and subject guides now makes it easier for patrons to find these resources by topic or title. Moreover, cross-referencing links to e-journals of a topic appear if a user searches for databases under that topic. This not only makes finding a database easier for patrons, it streamlines the maintenance of the Web site.

We were correct in adding an A to Z list in the initial redesign—it was heavily used during the user tests. A problem with the initial A to Z list that became apparent during the users tests and surveys was that it did not contain enough links. More than two hundred headings were added to the A to Z list, making it more comprehensive than before.

Our study provides a practical basis for further Web site studies, whether on the LSU Libraries' site, or other sites. To determine the final success of the study, further research is needed to determine whether ratings of the Web site and the success rate of searching for known items are higher than in our initial study.

Acknowledgements

The authors would like to thank Liz Shaw and Don King, both on faculty at the University of Pittsburgh, for taking the time to read and comment on the paper. Their suggestions made the paper stronger.

References

1. American Library Association, User Access to Services Committee, RUSA Machine-Assisted Reference Section, "User's Information-Seeking Behavior: What Are They Really Doing? A Bibliography," *Reference and User Services Quarterly* 40, no. 3 (spring 2001): 240–50.

2. Somporn Puttapithakporn, "Interface Design and User Problems and Errors: A Case Study of Novice Searchers," *RQ* 30, no. 2 (winter 1990): 195–204.

3. Tim Bucknall and Rikki Mangrum, "U-Search: A User Study of the CD-ROM Service at the University of North Carolina at Chapel Hill," *RQ* 31, no. 4 (summer 1992): 542–53.

4. Michael Culbertson, "Analysis of Searches by End-Users of Science and Engineering CD-ROM Databases in an Academic Library," *CD-ROM Professional* 5, no. 3 (Mar. 1992): 76–79.

5. Lara D. Catledge and James E. Pitkow, Characterizing Browsing Strategies in the World Wide Web. Accessed Apr. 1, 2002, www.igd.fhg.de/archive/1995_www95/proceedings/papers/80/userpatterns/UserPatterns.Paper4.formatted.html.

6. Laurel A. Carter, "Building a Better Reference Interface," *Internet Reference Services Quarterly* 3, no. 4 (1998): 57–84.

7. Jerilyn R. Veldof, Michael J. Prasse, and Victoria A. Mills, "Chauffeured by the User: Usability in the Electronic Library," *Journal of Library Administration* 26, no. 3/4 (1999): 115–40.

8. Jakob Nielsen, *Usability Engineering* (San Francisco: Morgan Kaufmann, 1994); R. Wise, The Three Click Rule, Dec. 15, 2000. Accessed Apr. 1, 2002, www.website-owner.com/articles/design/3clickrule.asp.

9. Jakob Nielsen, The Top Ten *New* Mistakes of Web Design, May 30, 1999. Accessed Apr. 1, 2002, www.useit.com/alertbox/990530.html.

10. Louis Rosenfeld and Peter Morville, *Information Architecture for the World Wide Web* (Cambridge, Mass.: O'Reilly, 1998).

11. David Robins, "Information Architecture, Organizations, and Records Management," *Records and Information Management Report* 17, no. 3 (2001): 1–14.

12. J. Dietrich, K. Gordon, and M. Wexler, Effects of Link Arrangement on Search Efficiency. Accessed Apr. 1, 2002, www.otal.umd.edu/SHORE/bs09/index.html#toc.

13. John Morkes and Jakob Nielsen, Concise, Scannable, and Objective: How to Write for the Web. Accessed Apr. 1, 2002, www.useit.com/papers/webwriting/writing.html.

14. Nielsen, *Usability Engineering*.

15. Human Factors Research Group, Ireland. Software Usability Measurement Inventory (SUMI). Accessed Apr. 1, 2002, www.ucc.ie/hfrg/questionnaires/sumi/index.html.

16. Rosenfeld and Morville, *Information Architecture for the World Wide Web*.

E2M: Automatic Generation of MARC-Formatted Metadata by Crawling E-Publications

Siew-Phek T. Su, Yu Long, and Daniel E. Cromwell

This paper presents a system called E-pub to MARC (E2M), which automatically generates MARC-formatted metadata by crawling e-publications. The functions of its two key components, the Web Crawler and the MARC Converter, are introduced. The paper presents the methods and tools used for building the system. The process of crawling and gathering pertinent metadata stored in the e-publications and the transformation of the metadata into MARC-formatted records are described in detail. The complexity of the crawling and the record generation processes are also described. A comparison between the cataloging process of e-publication using the computer-aided E2M process and manual cataloging is presented to illustrate that the E2M process is a more cost effective and efficient method of organizing and proving access to e-publications.

The proliferation of scholarly electronic publications (e-publications) on the Internet has posed a challenging problem for catalog librarians. The process of manually cataloging information in this medium is not only time consuming but also human-resource intensive. In light of this challenge, the impetus of the project team was to develop a more efficient and effective method to catalog this type of material with the aid of a computer.

Bibliographic control of Web resources and its related issues have been widely discussed and written about.[1] The issue of automating the e-publication cataloging process is an important one. However, little has been done in developing systems to automate the entire labor-intensive cataloging process using WebCrawler technology and techniques for automatic data conversion and loading. Although WebCrawlers have been used to extract information from Web pages, they are not programmed to extract the specific metadata needed for constructing catalog records and for loading them into bibliographic databases. For example, the two notable crawlers of the popular search engines, Google and Internet Archive, crawl the entire Web and extract keywords from Web pages to generate indexes for accessing relevant Web pages.[2] Meta-crawlers, such as MetaCrawler and Dogpile, integrate the search results obtained from different search engines.[3] Site-specific crawlers, such as WebSPHINX, allow users to specify site-specific crawling rules and perform so-called personal crawling.[4] The Hermes notification service system uses a component called wrapper to extract bibliographic data from HTML documents on publishers' Web sites and generate XML documents that contain bibliographic data.[5] The bibliographic data are typically the journal's table of contents (TOC). A commercially available tool for cataloging Web resources is the MARCit system.[6] The system provides a template for users to fill in such cataloging information as URL, author, title, and subject headings to convert the information to standard MARC-formatted records, which can be loaded into the local library management system. However, it does not have a WebCrawler component to automatically access and harvest the Web page metadata.

In contrast to the above systems, the E2M system described in this paper deals with the entire e-publication cataloging process. It starts with the automatic extraction of metadata from Web pages and goes on to the conversion of metadata into MARC-formatted records. Next, these records are loaded into the local system for authority verification. The final stage of the process consists of exporting verified MARC records to the Online Computer Library Center (OCLC) catalog for sharing with the bibliographic information community.

Project Domain

The e-publications housed in the Extension Digital Information Source (EDIS) database of the Institute of Food and Agricultural Sciences (IFAS) at the University of Florida (UF) is used as the project domain.[7] The EDIS database is the official electronic database of IFAS' current extension service and research publications. The reason for choosing the EDIS database as the project domain is a utilitarian one. It has been UF Library policy to catalog all IFAS publications. Currently there are more than five thousand documents in the EDIS database, out of which nearly four thousand are in electronic format.[8] Moreover, thirty to forty new e-publications are being added monthly. The benefits of automating the cataloging process for this increasingly large database are obvious. Furthermore, the structure of the IFAS e-publications is somewhat standardized, making it ideal to develop a WebCrawler to automatically harvest the meta-information.

Siew-Phek T. Su (pheksu@mail.uflib.ufl.edu) is Associate Chair for Central Bibliographic Services, George A. Smathers Libraries and **Yu Long** (yul@ufl.edu)is Graduate Student, Electrical and Computer Engineering Department, University of Florida, Gainesville; **Daniel E. Cromwell** (fcldec@nersp.nerdc.ufl.edu) is LMS Field Specialist, Technical Services, Florida Center for Library Automation, Gainesville.

Components and Operational Process of E2M

Figure 1 shows the components of the E2M system and its operational process.

The WebCrawler accesses and scans Web pages, extracting relevant metadata from them. The extracted metadata are represented in the form of data field names and values. They constitute the input to the MARC Converter. The converter transforms the metadata into MARC-formatted records, which are then loaded into the library's online management system. For this task, we use the FULOAD program.[9] The computer-generated records undergo an authority verification process before being loaded into the OCLC database as acceptable MARC records. We use the BatchBAM program as the authority verifier to do authority verification.[10] To transfer records to OCLC, we use the upload feature of the CLARR program developed by Gary L. Strawn of Northwestern University Library.[11] The FULOAD and the BatchBAM programs were also written by Strawn specifically to assist the UF library staff in the automated processing of bibliographic records. This paper describes only the two key components of the E2M system: the WebCrawler and the MARC Converter.

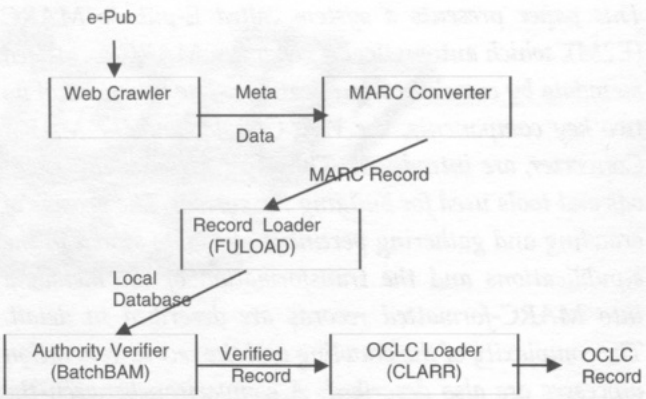


Figure 1: Components and Operational Process

WebCrawler and Metadata

What is a WebCrawler and why is it ideal to use it for harvesting metadata of e-publications? A WebCrawler is "a program that automatically traverses the Web's hypertext structure by retrieving a document, and recursively retrieving all documents that are referenced."[12] Crimmins presents an excellent review of Web crawling in which he examines various strategies and approaches used for developing different types of crawlers, such as scalable crawlers, agent-based crawlers, and crawlers that are designed for finding specific information. The design of a WebCrawler depends very much on its application.[13] As mentioned earlier, the WebCrawler developed for the E2M (E-pub to MARC) project is for finding specific metadata information in each e-publication page, such as author, title, publisher, date of publication, and notes necessary to generate a MARC record.

The e-publications residing in the EDIS database have a somewhat standardized structure. Each e-publication has all or part of the following metadata: titles, author, section titles, subsection titles, summary, bibliography, footnotes, and copyright information, in that order. However, the specific format and the placement of these metadata vary in different e-publications. The challenge for the crawler is to be able to detect these differences by scanning the HTML representation of each e-publication in which these data fields are not explicitly tagged. This, of course, is more difficult than scanning and extracting metadata from an XML-formatted document in which metadata are tagged by field names. Our solution is to make the best use of the tags in HTML to find the pertinent metadata we need to construct the bibliographic record in MARC format. The following examples illustrate the challenge and the solution:

The title of the e-publication always appears as the first data element in an IFAS e-publication. For authors, which usually come after the title, we search for the information between the tag "</h1>", which is the end tag for the section containing the title information, and the tag "<sup>", which is the beginning tag for the superscripted footnote number that always follows the last author's name.

A publication number (PN) is located in the footnote section of the e-publication and has the following possible formats:

- Alpha-numeric string with the alpha part always in the upper case, such as:
  - NEY-250
  - FCS 8155
- Name followed by number, such as:
  - Bulletin 810
  - Fact Sheet ENH-88

Using the same method for locating the author information, the program looks for the feature of the PN, starting with a letter in upper case and ending with a numeral.

For the publication date, which also resides in the footnote section, the program searches for the four-digit year number.

In the process of harvesting the relevant information from e-publications, the crawler occasionally encounters

errors such as incorrect URLs, broken URLs, incorrect HTML tags, or some type of system-generated errors. When an error occurs, the crawler records the error in an error file. The search and extraction process will continue, leaving those errant e-publications to be dealt with individually by the staff. Thus, the process will not be blocked when a noted error is encountered.

The Crawler uses a breadth-first-search algorithm to locate all linked pages.[14] The search starts from a root page, fetches all the URLs on that page, and puts them into a first-in-first-out (FIFO) queue. Next, it uses the first URL in the queue to locate the next page, extracts all its URLs, and puts these newly fetched URLs into the queue again. Whenever it finds a URL for an e-publication, it starts to extract the relevant metadata. It is called a breadth-first-search method because it will finish searching all the URLs on one level before searching for the URLs on the next level. To illustrate the algorithm more clearly, we provide the flowchart of the crawling process in figure 2.

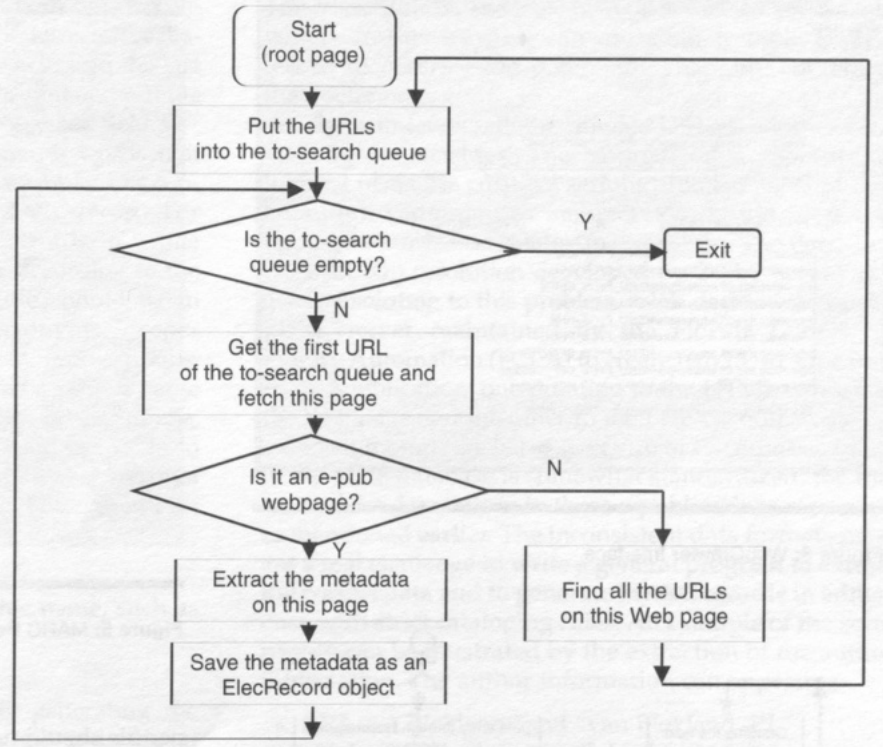


Figure 2: Crawling Process

The user enters the starting URL of a Web page (i.e., the root page) through the WebCrawler interface (shown in figure 3). The system implementation and user interfaces are described in detail in the later part of this paper.

MARC Conversion Process

The MARC conversion process deals with the automatic generation of a standard MARC-formatted record based on the extracted metadata of an e-publication that is embedded in the crawling process part illustrated in figure 4. In the conversion process, two types of data need to be combined to form a MARC record, the crawled data and the constant data. The crawled data are extracted from e-publications. Data, such as author and title, vary from e-publication to e-publication. The constant data are data that should appear in all generated MARC records. Some examples of constant data are given below:

1. Publisher information in the 260 field: "[Gainesville, Fla.] : University of Florida Cooperative Extension Service, Institute of Food and Agricultural Sciences, EDIS"
2. Technical note in the 538 field: "Internet access required."
3. At head of title in the 500 field: "University of Florida, Cooperative Extension Service, Institute of Food and Agricultural Sciences, EDIS."

The constant data consists of two subtypes: the same data that appear in every IFAS e-publication (see examples 1 and 3), and the data that are necessary to form valid MARC records according to cataloging rules (see example 2). The user input the constant data through the RecordFormat interface shown in figure 5.

The crawled data are stored in memory as an ElecRecord object and the constant data are stored as a MarcFormat object. These data are used to construct a MARC-formatted record based on the structural specification of the MARC bibliographic record. In order to achieve better performance, we store the generated MARC record in an in-memory buffer called OutputRecord instead of writing it to a file immediately after its generation. The accumulated MARC records in the buffer are written to a file only when the buffer is full. After the output operation,

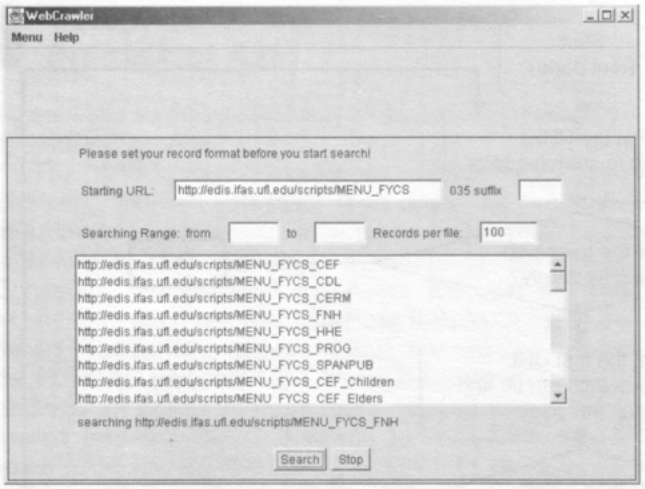


Figure 3: WebCrawler Interface

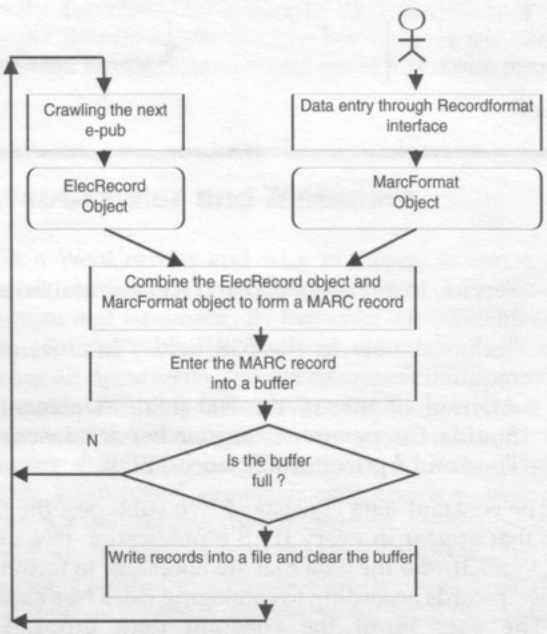


Figure 4: Conversion Process

the buffer is cleared and reused for the next batch of records. Each batch contains the number of records set by the user through the WebCrawler user interface as the value of the variable named "Records per file" (see figure 3). In this way, we can reduce the number of time-consuming output operations. However, the value of Records

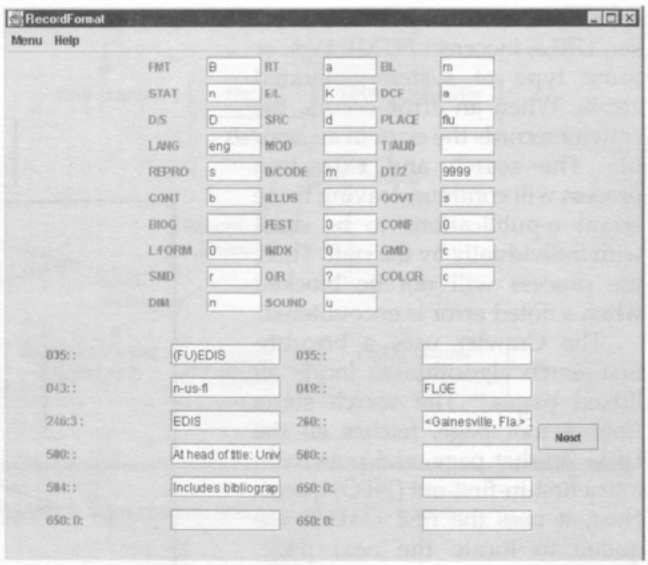


Figure 5: MARC Record Form

per file should not be set too high because some computers that run the E2M system may have a limited amount of main memory space. From our experience, we recommend that the value should not be set higher than one hundred records.

For creating a record with the proper MARC format, we first determine the various data fields and values that are to appear in the finished product. We then create a field-by-field list of cataloging rules to be used by the converter in placing data fields and values in the proper format of a valid MARC record. The generated record can then be confidently used by the bibliographic community. We base the rules on the MARC format standard specification, the Anglo-American Cataloguing rules (AACR2), the International Standard for Bibliographic Description (ISBD[ER]) and OCLC Bibliographic Formats and Standards.[15] For example, the rule for the author/title field is provide only one author access point and observe the following rules:

- Use the first-named author as the main entry (100:1)
- If more than three authors, use title as the main entry (245:0) and the first-named author as an added entry (700:1)

In addition to these rules, we have to introduce additional rules for handling metadata elements that do not conform to the data field restriction of the local online library system. For example, the 520 field of the NOTIS system has a length restriction of one thousand characters. However, the summary of an e-publication may

exceed this limit. It does not make sense to simply truncate the summary after the thousandth character. A reasonable rule for the converter to follow is to find the last sentence that fits the thousand-character limit and put the remainder of the summary in an additional 520 field(s).

The upper and lower cases of letters in words that appear in an IFAS e-publication pose a problem in converting them to the correct cases in a MARC record. For example, in an IFAS e-publication, all words in a title begin with uppercase letters. However, according to the AACR2, only the first word in the title should be in uppercase, with the exception of acronyms, proper nouns, and directional words followed by proper nouns. To deal with this problem, we developed a proper name table, which consists of a set of acronyms, proper nouns, and directional words. The converter uses this table to identify those words whose uppercase letters should not be changed in the process of conversion. Some examples of words in the table include:

- geographic names, such as "Florida";
- directional word before a geographic name, such as "South Florida"; and
- acronyms, such as "EDIS."

Before the process of automatically generating the MARC record could take place, we also had to decide on the type of record we would like the converter to generate. We opted for encoding level K (less-than-full level) records to avoid having to assign subject headings and call numbers, and to establish entries for secondary authors. To make up for the lack of subject headings, we decided to include section and sub-section titles in content notes and the summary as summary notes to provide keyword access.

Another problem we had to solve was how to handle different versions of an IFAS e-publication. Unlike printed documents, for which each edition is separately published and cataloged, only the current version of an IFAS e-publication is made available, replacing the old version. The new version has the same URL as the old version and has a notation in the text indicating that the publication date is the revision date. The policy of keeping only the current version creates a problem and a challenge for maintaining the accuracy of a bibliographic record. Technically, an existing bibliographic record should be modified to contain the publication date of the new version. However, it is quite costly not only to update the date each time a new version appears but also to track when the new version appears. The approach we have taken to deal with the current version problem is to code the type of date and publication status as "m" (a range of dates) and leave an open-ended publication date. We also include a note to indicate the date that the crawler viewed the Web page. An example of this is "Title from Web page viewed on July 25, 2001." The view date is the date that the metadata was harvested. In this way, we can use a single bibliographic record to describe the potentially changing content of the document.

The maintenance of the crawled URLs is another issue we had to consider. The unpredictable mobility of Internet resources creates a serious problem for librarians because it compromises their services to the users and imposes a burden on catalog maintenance. The Persistent URL (PURL) resolution developed by OCLC serves as a general solution to this problem.[16] We decided to use the PURL server maintained by the Florida Center for Library Automation (FCLA) to create two PURLs for each IFAS e-publication; one pointing to the HTML version of the Web page and the other to the PDF version.

Even though the basic structure of the e-publications in the EDIS database is somewhat standardized, the formats of the data values in these e-publications may vary, as mentioned earlier. The inconsistent data formats present a real challenge to write a general program to extract the correct data and to generate MARC records in adherence with strict cataloging rules. An example of the complexity can be illustrated by the extraction of the author information. The author information can appear as:

- "P.J. van Blokland" and "van Blokland, P.J."
- "John Smith, Ph.D." or "John Smith, Jr."
- two or more authors separated by "and," by comma, or by space

In the first example, it is not possible for the program to distinguish correctly which character string constitutes the first name and which the last name. In the second example, unless a specific rule is written for the program to ignore titles such as "Ph.D." and "Jr.," it will not know that these characters are not a part of the name. In the third example, there is no way for the program to count the number of authors when there is not a consistent way of separating each author unless all the possible separators are made known to the program.

Our current solution is to identify as many different patterns and structures of metadata as possible and to program the converter to recognize them. However, this approach requires that the program code be modified or extended if more deviant patterns and structures are found. A better solution is to introduce a rule language and write the deviant patterns and structures as rules. The converter can then be programmed to interpret the rules during the parsing of the character string to correctly extract the metadata. The development of such a rule-driven parser for the converter is contemplated.

There are a number of Spanish language e-publications in the EDIS database. The use of diacritics in the Spanish language text posed a unique challenge. The representations of such diacritics as acute, tilde, and umlaut in the HTML source code are different from the MARC 21

Specification for Character Sets needed to construct MARC records.[17] We have programmed the system to perform the conversion of acute, tilde and umlaut, which are the most common diacritics in the Spanish language. Future work can be done to set up a lookup table for special characters in different languages.

The screen-shots in figures 6, 6b, 7, and 8 serve to summarize the transformation process starting from an e-publication to MARC communication format record to the final product of the OCLC record. Figure 6 and 6b show the Web pages of an IFAS e-publication that the WebCrawler extracted; the image has been split to show the beginning and ending of the document. The complete document can be found at the following Web site: http://edis.ifas.ufl.edu/HE796 (Accessed April 30, 2002).[18] Figure 7 shows the corresponding harvested record in MARC communication format and figure 8 shows the corresponding OCLC MARC record.

The transformation of the author's name can be seen as the record evolved. As shown in figure 6, the form of the author's name in the e-publication Web page is "Garret D. Evans, Psy.D." The harvested MARC record in figure 7 shows that the form of the name has been transformed to "Evans, Garret D." with the elimination of the title, "Psy.D." The final product, as shown in figure 8, presents the authoritative form of the name of the author, "Evans, Garrett D, 1965– ," after the record has gone through the authority verification process. A cataloger manually changed the form of the name to the established one. An NAR (Name Authority Record) is created if the correct form has not been established. The number of our NAR submissions to NACO (Name Authority Cooperative) Program has increased due to the loading of these E2M-generated records. Another example to illustrate the transformation of the e-publication to the MARC record is in the title proper. In figure 6, the title displays as "The 'Fool-Proof' Time-Out." E2M transformed the title into "The 'fool-proof' time-out," with the proper lower-case letters as shown in figures 7 and 8. The second 500-field, with the quoted note as shown in figure 8 and its corresponding MARC communication record in figure 7, is extracted from the Footnotes section of the corresponding e-publication ending page shown in figure 6: "This document is Fact Sheet FCS 2113, a series of the Department of Family, Youth, and Community Sciences, Florida Cooperative Extension Service, Institute of Food and Agricultural Sciences, University of Florida. Publication date: April 1997. First published as HE 2101 June 1996. Reviewed: April 1997."

From the Footnote information, the WebCrawler also harvested pertinent data such as the publication date (260 |c) and the publication number (246 |b). Additional information in other fields and subfields, such as the 043, 246 (except for the publication number), 260 (except for the publication date), 500 "At head of title" note, and 538

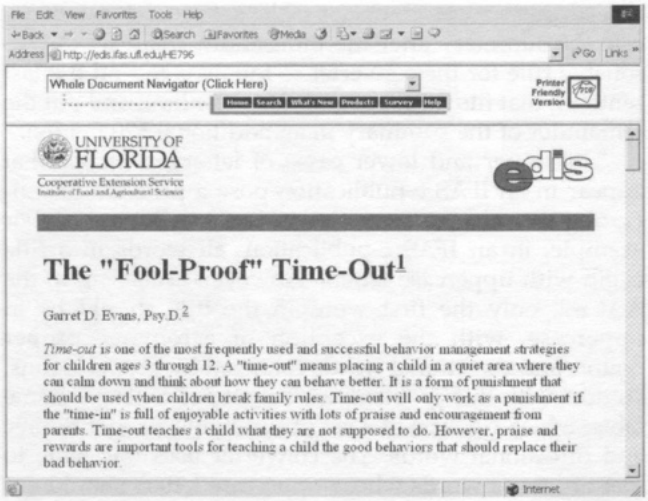


Figure 6: E-Publication Web Page, Beginning

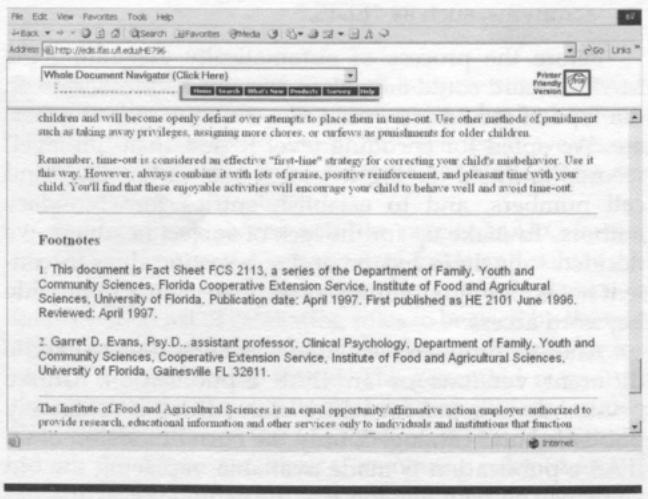


Figure 6b: E-Publication Web Page, End

fields, are the constant data taken from the record template, as shown figure 5.

System Implementation and User Interfaces

The key components of the E2M system, the WebCrawler and the MARC Converter, are both written in the Java programming language.[19] We chose Java not only because

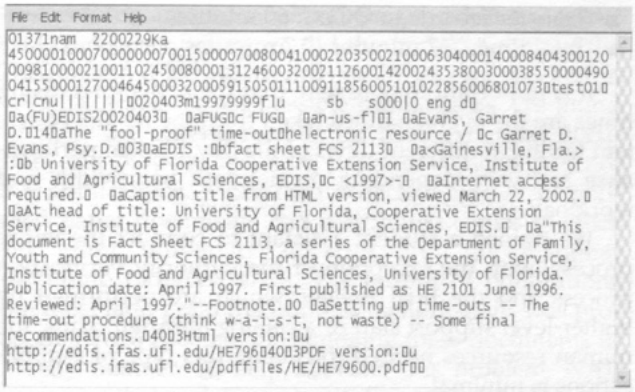


Figure 7: MARC Communication Format

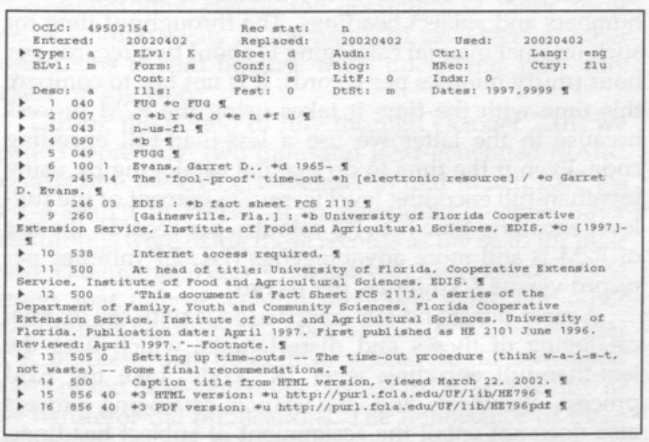


Figure 8: OCLC Record

its platform-independent feature allows us to install the system on any Java-enabled computer, but also because its object-oriented feature is easy to program, debug, and update.

To crawl for metadata of e-publications, different library institutions or users may want to extract different metadata and convert them into the desired MARC records. To avoid reprogramming the converter each time a different MARC record is desired, we use a form-driven approach to implement the converter. A record template that contains a standard set of data fields for monographs is predefined and accessed as a form by the user through a browser (see figure 5). The user completes the form by assigning values to the data fields shown in the template. All these values will appear in each of the generated MARC records. For example, in figure 5, the Format (FMT) field will always be coded as "B," the Record Status (STAT) coded as "n," and the Encoding Level (E/L) as "K" to indicate that the generated MARC record is a less-than-full level bibliographic record.

After filling in the first form, the user accesses the second form shown in figure 3. This form allows the user to enter the URL of a Web page (i.e., the root page), from which the Crawler would do a breadth-first search for all the hyperlinks as described in the WebCrawler and Metadata section of this paper. The user can also enter the URL of a particular e-publication for accessing the document directly.

The user interface shown in figure 3 also allows the user to specify a range of dates for traversing a subset of hyperlinks that fall into the specified range, or a set of specific hyperlinks for accessing the e-publications pointed to by them. For example, in the Web page that contains the hyperlinks to EDIS' new documents, the hyperlinks are partitioned by dates as shown in figure 9. The user may want the crawler to crawl for hyperlinks within the date range from September 21, 2001, to September 14, 2001. (Note that the dates are reversed because of their position on the Web page. See figure 9.) Another option the user has is to enumerate those hyperlinks from which the crawler should perform the search (e.g., IG148 to AN110). The reason for providing these two search options is to give the user the added flexibility to specify which e-publications are to be processed.

The "Records per file" field in figure 3 is for setting the desired number of records for each output file. The "035 suffix" field is for inputting a unique suffix to be included in the 035-field (system control number field).

To facilitate the editing of the proper name table, we developed a form consisting of three function-buttons: "Add," "Find," and "Delete" for adding, finding, and deleting entries from the table, respectively (see figure 10).

Discussions and Ongoing Work

We implemented and fully tested the system using a sample of approximately five hundred e-publications. Experimental results indicated that the system was an effective and efficient method to catalog the e-publications residing in the EDIS database. We then put this system to use in the real library environment in July 2001. Using the E2M system, we generated and loaded over 2,500 MARC-formatted records in our local online database and uploaded more than two thousand records to OCLC. The process of harvesting the relevant metadata of the e-publications, converting the crawled metadata to the MARC-formatted records, and loading the MARC records into the local database was very fast. The most time-consuming part was the manual review of each

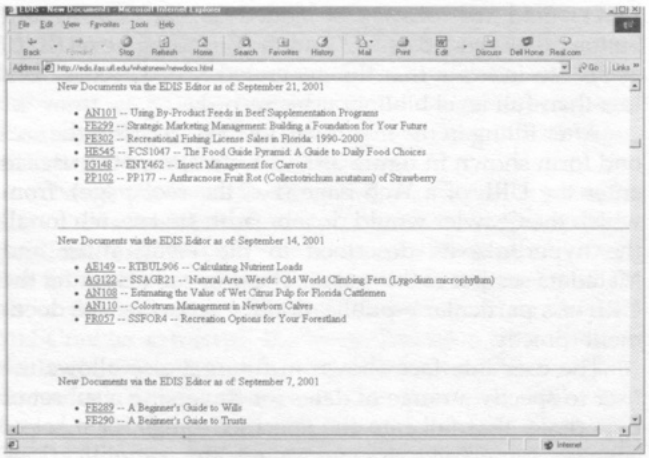


Figure 9: EDIS New Documents Screen

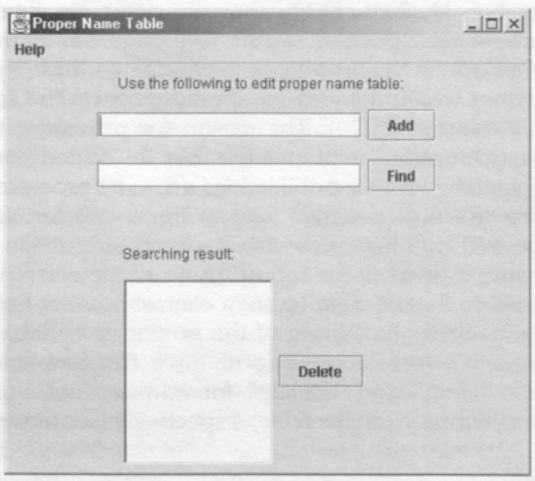


Figure 10: Proper Name Table

record and the follow-up work resulting from the authority verification process to make sure that the records met acceptable standards for sharing with the bibliographic universe. The individual tasks and their throughput time per hundred records were:

- Crawl for metadata and convert them to MARC records: 10 minutes
- Load records into online library management system: 2 minutes
- Authority verification and corrections: 180 minutes
- Create PURLs: 15 minutes
- Transfer records to OCLC: 60 minutes
- Total time: 267 minutes (2.7 minutes per record)

It is important to note here that the above processing times are the average times taken from the results of system testing at different times of the day. System performance can vary depending on network bandwidth and workload.

We trained a student assistant to conduct the entire process, with the exception of the authority verification process and the final review, which are performed by higher-level support staff or a catalog librarian. Thus, the human resources needed to catalog this type of e-publications is minimal.

A comparison between the cataloging process of IFAS e-publications using the computer-aided E2M process and manual cataloging would certainly be useful. When we cataloged the IFAS publications in the past, we used the full-level encoding code, including assigning call numbers and subject headings. The throughput time for such original manual cataloging is about two records per hour (thirty minutes per record). It is not fair to compare this time with the time it takes using the E2M process because in the latter we use a less-than-full encoding code. Even if the time to do manual cataloging by using less-than-full encoding level is, say, 25 percent of the full-level encoding time (i.e., 7.5 minutes per record), the use of E2M is still more advantageous (i.e., 2.7 minutes per record versus 7.5 minutes per record).

A more direct and fair comparison is with the manual cataloging of theses and dissertations, which uses the less-than-full encoding level "K" code. Like the E2M process, the thesis and dissertation cataloging process also does not entail the assignment of subject headings, and several fields in the record also have constant data. Its authority verification process and the record transfer process to OCLC are quite similar to those of the E2M process. A student assistant is also involved in the theses and dissertation cataloging process. Our performance measure for cataloging theses and dissertations manually is about 7 records per hour (i.e., 8.57 minutes per record) as compared with the 2.7 minutes per record cataloging IFAS e-publications using the E2M process. It is obvious that the E2M cataloging process is more efficient and effective, thus a less costly method of cataloging this type of e-publication. A side benefit for using the E2M system is that it allows us to contribute original bibliographic records to the OCLC database and gain monetary credit of $4.05 per record for the library.

All the E2M-generated records loaded to OCLC are standard MARC less-than-full encoding level records that follow AACR2. To avoid loading records with errors, these harvested records went through manual review as well as through an authority verification process in which changes were made where necessary.

These records are also new and unique additions to the OCLC database. Before we did the initial loading of the E2M-generated MARC records to OCLC, we checked in the OCLC database for possible duplications. We found that there are several records for IFAS extension documents, but they are for older paper documents and not for e-publications. More than a decade ago, IFAS published paper documents and distributed copies to other land grant university libraries and the National Agriculture Library (NAL). The output of publications from most land grant institutions, including UF, was prolific, causing a problem of cataloging these documents in a timely manner. To help alleviate this problem, NAL encourages that each home institution takes the responsibility for cataloging its own publications. Thus, it is unlikely that a library will catalog other institutions' extension publications. In the meantime, IFAS abandoned publishing extension documents in paper format in favor of e-publications. Hence, the possibility of these documents landing easily in the hands of catalogers at other institutions is slim because these documents now only reside on the Web.

During the period of July through October 2001, we loaded more than two thousand E2M-generated MARC records into the OCLC database. In mid-March 2002 we did a search in WorldCat to determine whether other institutions were using these records as the basis for their own cataloging. As we suspected, the search result shows that only a handful of records have one other institution's holdings symbol besides UF. We believe that it is unlikely that other institutions would download many of the E2M-generated records in OCLC because of the nature of the publications. The importance of these records going into the OCLC database is that they are now available on FirstSearch and WorldCat to be used by researchers, not so much to be downloaded into other library databases. From our many years of experience in cataloging extension publications in OCLC, we have found that there has not been a pattern of libraries aggressively seeking to catalog other institutions' electronic extension documents, except selectively. We also believe that, in general, if other libraries were to derive our E2M-generated MARC records, some of them may improve these records by adding subject headings and perhaps classification numbers. Otherwise, the standard description that we provided as a less-than-full encoding level record would be sufficient. This proves to be true from the search we did in mid-March 2002. We checked the library online database of a deriving institution and found that subject headings have been added locally to the derived record. None of the records that have other institutions' holdings symbols attached have been enhanced in OCLC. It would be interesting to track the usage of these records in the OCLC and WorldCat databases by doing a periodic search similar to the one conducted in March 2002 to determine if the initial analysis remains the same. There is a possibility that the result of the analysis would be different due to the aging of these records, an increased chance that these records would be found and used.

In addition to the AACR2 MARC-quality records, and the cost-effective and efficient method of organizing and providing access to the IFAS e-publications, the E2M project demonstrates two of the action steps identified by the Library of Congress in its "Bibliographic Control of Web Resources: A Library of Congress Action Plan."[20] The E2M system is the product of a collaborative effort among the IFAS Information Technology group, the researchers in the Database Systems Research and Development Center of the Computer and Information Science and Engineering Department, and the UF librarians. The collaboration leads to the ongoing exploration of better methods for organizing Web resources and development of other information access and dissemination mechanisms. This project has developed and demonstrated WebCrawler technology, techniques for automatic metadata extraction from e-publications and automatic generation, and loading of standard MARC-formatted records into the local library online management system and the OCLC database. The computer-generated records include such additional descriptive information as tables of contents and summaries to enable keyword searching. The records also provide for explicit linking from the bibliographic record to the e-publication via two PURLs.

Although the E2M project is a work in progress, it has already proven to be effective in automatic cataloging. A few tasks can be done to further improve the performance and flexibility of the system. For example, batch transfer to OCLC, instead of record by record, will improve the data transfer time. The creation of PURLs can be automated and programmed as a part of the WebCrawler's function. Our ongoing work also includes developing a rule-driven parser for the crawler so that it can extract metadata from other types of e-publications without having to modify or reprogram the crawler. A look-up table for special characters in different languages will allow the system to process non-English e-publications. All these tasks are part of our ongoing work to make E2M a more generic software system.

Acknowledgements

We would like to thank Jimmie Lundgren, Betsy Simpson, Diana Hagan, and Howard Beck for their helpful comments and valuable discussions on this project. We would also like to thank Professor Stanley Su for his support and advice on the technical aspect of this project.

References

1. Abdus Sattar Chaudhry, "A Study of Current Practices of Selected Libraries in Cataloging Electronic Journals," *Library Review* 50, no. 9 (Oct. 2001): 434–43. Martin Dillon and Erik Jul, "Cataloging Internet Resources: The Convergence of Libraries and Internet Resources," *Cataloging and Classification Quarterly* 22, no. 3/4 (1996): 197–238. International Federation of Library Associations and Institutions, Functional Requirements for Bibliographic Records. Accessed Feb. 27, 2002, www.ifla.org/VIIs13/frbr/frbr.pdf, March 1998. Kristen H. Gerhard, "Cataloging Internet Resources: Practical Issues and Concern," *The Serials Librarian* 32, no. 1/2 (1997): 123–37. Yiu-On Li, "Computer Cataloging of Electronic Journals in Unstable Aggregator Database: The Hong Kong Baptist University Library Experience," *Library Resources and Technical Services* 45, no. 4 (Oct. 2001): 198–211. Proceedings of the OCLC Internet Cataloging Colloquium, San Antonio, Jan. 19, 1996. Accessed Feb. 27, 2002, www.oclc.org/oclc/man/colloq/toc.htm. University of Virginia, "Proceedings of the Seminar on Cataloging Digital Documents." Accessed Feb. 27, 2002, http://lcweb.loc.gov/catdir/semdigdocs/seminar.html. "Tools for Cataloging Internet Resources." Accessed Feb. 27, 2002, www.bowdoin.edu/~kfattig/netsl.

2. Google. Accessed Dec. 10, 2001, www.google.com. Internet Archive. Accessed Dec. 10, 2001, www.archive.org.

3. Metacrawler. Accessed Dec. 10, 2001, www.metacrawler.com. Dogpile. Accessed Dec. 10, 2001, www.dogpile.com.

4. WebSPHINX: A Personal, Customizable WebCrawler. Accessed Dec. 10, 2001, www.cs.cmu.edu/~rcm/websphinx.

5. D. Faensen et al., Hermes—A Notification Service for Digital Libraries. Accessed Sept. 30, 2002, www.inf.fu-berlin.de/inst/ag-db/publications/2001/jcdl01.pdf.

6. MARCit. Accessed Feb. 27, 2002, www.marcit.com.

7. EDIS Web Site. Accessed Dec. 10, 2001, www.ifas.ufl.edu.

8. Information on the EDIS Database. Accessed Dec. 10, 2001, http://it.ifas.ufl.edu/newsletter/edis.html.

9. Gary L. Strawn, "Instructions to Accompany FULOAD.EXE: A Record Management Program Written for the University of Florida." Accessed Dec. 10, 2001, http://web.uflib.ufl.edu/rs/fuload/home.html.

10. Gary L. Strawn, "Batch BAMming: A Method for the Automatic Validation of Headings in Defined Sets of Bibliographic Records." Accessed Dec. 10, 2001, http://web.uflib.ufl.edu/rs/rsd/batchbam.html.

11. Gary L. Strawn, "User's Guide to Accompany CLARR, the Cataloger's Toolkit." Accessed Dec. 10, 2001, http://staffweb.library.northwestern.edu/catalog/clarr/home.html.

12. M. Koster, "The Web Robots Pages." Accessed Dec. 10, 2001,http://info.WebCrawler.com/mak/projects/robots/robots.html.

13. Francis Crimmins, "WebCrawler Review." Accessed Dec. 10, 2001, http://dev.funnelback.com/crawler-review.html

14. M. Najork and J. Wiener, "Breadth-First Search Crawling Yields High-Quality Pages," in *Proceedings of the Tenth International World Wide Web Conference*, Hong Kong, May 2001. Accessed Dec. 10, 2001, www10.org/cdrom/papers/208/index.html.

15. Library of Congress, MARC Standards. Accessed Dec. 10, 2001, www.loc.gov/marc/marc.html. American Library Association, *Anglo American Cataloguing Rules*, 2d ed. 1998 rev. (Chicago: ALA, 1998). International Federation of Library Associations and Institutions, ISBD(ER): International Standard Bibliographic Description for Electronic Resources. Accessed Dec. 10, 2001, www.ifla.org/VII/s13/pubs/isbd.htm. OCLC, *Bibliographic Formats and Standards*, 2d ed. (Dublin, Ohio: Online Computer Library Center, Inc., 1996).

16. Keith Shafer, Stuart Weibel, and Erik Jul, Introduction to Persistent Uniform Resource Locators. Accessed Dec. 10, 2001, http://purl.oclc.org/OCLC/PUR L/INET96.

17. MARC 21 Specifications for Record Structure, Character Sets, and Exchange Media CHARACTER SETS: Part 3, 2000. Accessed Dec. 10, 2001, http://lcweb.loc.gov/marc/specifications/specchartables.html.

18. Garret D. Evans, The "Fool-Proof" Time-Out. Accessed Apr. 30, 2002, http://edis.ifas.ufl.edu/HE796.

19. Thom Blum et al., Writing a WebCrawler in the Java Programming Language. Accessed Jan. 23, 2002, http://developer.java.sun.com/developer/technicalArticles/ThirdParty/WebCrawler.

20. Library of Congress, Bibliographic Control of Web Resources: A Library of Congress Action Plan. Accessed Feb. 27, 2002, www.loc.gov/catdir/bibcontrol/actionplan.html.

Library Systems and Unicode: A Review of the Current State of Development

Laura Tull

Unicode, a standard developed in 1991, defines a universal character set for encoding the characters in the scripts of the world's languages. Unicode implementation has been gaining momentum in recent years especially in the software and computer industry. Academic libraries with collections of materials in multiple languages will want to take advantage of Unicode for display and searching of materials in non-Latin scripts such as Arabic, Hebrew, and Chinese. The focus of this article is a review of Unicode and its incorporation in library systems.

Unicode is a standard for a universal character set for encoding the characters in the scripts of the world's languages. It is a fully compatible 16-bit version of an international standard developed by the International Organization for Standardization (ISO) and the International Electrotechnical Committee (IEC), known as ISO/IEC 10646 or the Universal Multiple-Octet Coded Character Set. The Unicode standard was developed by a consortium of interested parties, known as the Unicode Consortium, composed mainly of such computer industry giants as IBM, Apple Computer, Adobe Systems, and Microsoft. However, the library world also had a stake in its development. Research Libraries Group (RLG), Online Computer Library Center (OCLC), and several library system companies are members of the consortium.

Although the earliest version of the Unicode standard was published in 1991, it is just beginning to be fully incorporated into many systems. Its use is prevalent in the computer industry in such programming languages as Sun Microsystems's Java and such operating systems as Microsoft Windows NT and Apple Computer's Mac OS 8.5. Popular Web browsers, including Microsoft Internet Explorer, Netscape Navigator, and the latest version of Opera also support Unicode. With Web browsers that can be set to view Unicode, switching back and forth between different character sets is no longer necessary. Taking advantage of the Unicode character set has become easier in recent years with the availability of Unicode fonts, such as Code 2000 and Code 2001, a shareware font produced by James Kass, and Microsoft's Arial Unicode MS.[1]

This standard is important to libraries that collect materials in many different languages and want to be able to display the native scripts in their Web catalogs as well as allow users to search by typing in the native scripts. It is especially useful when trying to display a record that may have multiple scripts, such as a record that may contain both Arabic and Hebrew. The idea behind Unicode was to develop one international character set for all of the scripts of the world's languages. One unique code would represent each character, even if that character were used in multiple languages. This could replace the multiple older and sometimes incompatible character sets that are presently in use, including the American Standard Code for Information Interchange (ASCII) and the Extended Binary Coded Decimal Interchange Code (EBCDIC) used by the computer industry; ISO 8859 character sets, such as Latin 1; and other character sets used in MARC records, such as the East Asian Character Code (EACC) set. Unlike most of the older 7- and 8-bit character sets, which were limited to 256 characters or less, the Unicode standard is based on 16 bits, allowing more than 65,000 characters to be encoded.

When it comes to transforming Unicode characters into bits and bytes that can be stored or transmitted by computer systems, there are three encoding forms that can be used to comply with the standard: UTF-8, UTF-16, and UTF-32. UTF-8 uses from one to four 8-bit code units or bytes. The big advantage of UTF-8 is that ASCII and its equivalent Unicode characters have the same value, making it more compatible with existing software. Most Web browsers also support UTF-8. UTF-16 uses one to two 16-bit code units, and UTF-32 uses a single 32-bit code unit. UTF-16, with its use of two 16-bit code units, allows for the encoding of more than a million characters.

A crucial issue before libraries is how Unicode implementations will affect the exchange of records. Libraries import records into their library systems from bibliographic utilities, such as OCLC and RLG, that have huge databases of bibliographic records representing scripts in all languages. Record exchange between libraries as well as with all sorts of companies around the world is now quite common, and library systems will eventually need to be able to import and export records in Unicode to do business. During the interim period when everyone is still implementing Unicode, companies will have to offer record exchange in both MARC 21 character sets and Unicode until Unicode becomes the dominant character set used by most systems. The U.S. MARC Advisory Committee approved a proposal to allow bibliographic records to be encoded using Unicode in 1998. Currently, the encoding of MARC records is limited to UTF-8 by MARC 21 Specifications for Record Structure, Character Sets, and Exchange Media. The characters that can be encoded in UTF-8 are also currently restricted to the MARC 21 repertoire of characters for the purpose of enabling round-trip record exchange during the transitional period when systems are switching to Unicode. There has been much discussion about these restrictions.

Laura Tull (tull.9@osu.edu) is Systems Librarian at The Ohio State University Libraries, Columbus.

Some people are anxious to take advantage of the full Unicode character set, which contains many more characters than MARC. Questions have arisen about how we will define when the period of transition is over so that we can ease these restrictions. There has also been discussion about expanding the MARC 21 repertoire in order to be able to use more characters. At this point, we still need to proceed with caution and at the same pace to make sure that all parties engaging in record exchange are in sync. Otherwise, data could be lost in the process.

Literature review

A review of the library literature reveals many articles explaining the Unicode standard but few about its implementation. Erickson provides a good background for the history, purpose, and details of this standard and delves into the history of character sets commonly used in libraries.[2]

Zhang and Zeng speak about many of the barriers to adopting and implementing the Unicode standard in libraries, especially in East Asia.[3] Some of the barriers they specifically mention are that Unicode was not supported in the MARC standards; major bibliographic utilities such as OCLC and RLG did not have plans to implement Unicode; and only a few library system vendors were experimenting with implementing the standard. They note that the new Unicode standard would require much development by various sectors of the library world. They also mention the concern of East Asian librarians that the number of Han characters (used in Chinese, Japanese, and Korean) in the Unicode standard is not enough to support characters that occur in personal names, full-text documents, and rare books.

Aliprand provides background information on the Unicode standard and addresses some of the cataloging issues in libraries.[4] She iterates the design principles of Unicode, focusing in particular on the principle of encoding characters, not glyphs. In other words, a character is encoded only once even though it may appear in several forms when written. Defending this principle against those who advocate exact transcription of a work, she points out that catalogers have always made compromises when transcribing works that contain unique characters and mathematical symbols. The important issue to focus on is how a user will be able to retrieve this information. Encoding characters instead of glyphs is more efficient, less complex, allows for more effective searching by users, and requires less software development.

Chachra describes some of the technical complexities that arise in a Unicode implementation as well as other issues that need to be addressed.[5] He describes the elements of a true Unicode implementation, but makes the point that Unicode is just one aspect of support for languages in a library system. Other aspects include creating language translation tools, developing sorting mechanisms, dealing with word parsing in specific languages, and developing character mapping tools during the transition phase from the older character sets.

Many of the obstacles mentioned by Zhang and Zeng have been addressed in recent years.[6] Unicode was approved for encoding MARC records in 1998. The MARBI Character Set Subcommittee started to develop mappings to Unicode from most of the MARC character sets in 1996 but, being unfamiliar with Chinese, Japanese, and Korean, they established the East Asian Character Set Task Force to deal specifically with mappings from EACC to Unicode. The task force reviewed character mappings that the Unicode Consortium had already completed and identified missing characters. They also dealt with mappings for Korean Hangul and Japanese kana in addition to punctuation and component characters. This work, resulting in MARBI Proposal no. 2001-09, was completed and approved in 2001. Now that their work is complete, companies that are converting to Unicode can start to incorporate some of these changes.

OCLC and RLG have also started to work with Unicode. RLG started using Unicode in 2000 in Eureka, the Web-based interface to their databases. The records are currently stored in the MARC 21 character sets but converted on the fly to UTF-8, an encoding format that complies with the Unicode standard, for display in an appropriate Web browser. Because RLG has already developed the programs to convert records to UTF-8, they are positioned to convert MARC 21 records in the future as demand arises. In summer 2001, OCLC announced that it had selected the Oracle database technology to replace its aging proprietary system for WorldCat, its bibliographic database. This will allow the organization to support Unicode and eventually encode and store the vast amount of records representing materials in languages from all over the world in Unicode. Although they will be operating dual systems for some time, this won't continue indefinitely; those libraries that import records from OCLC will need to have library systems in place that can accept those records.

One of the obstacles mentioned by Zhang and Zeng was the development required by companies that produce library systems.[7] A survey of library system companies to see how this development is progressing is the focus of this article.

Method

The author's goal was to survey library system vendors who marketed their system to academic libraries to assess

where this segment of the industry stands in relation to Unicode implementation. The library system industry is small, competitive, and, with mergers and buyouts, seems to shrink regularly. The author identified companies by using information from the survey of library system vendors that appears annually in *Library Systems Newsletter,* published by the American Library Association.[8] The author also used a Web site, Library Technology Guides: Key Resources and Content Related to Library Automation, created by Marshall Breeding at Vanderbilt University.[9] If it was still unclear whether or not a company marketed their system to academic libraries, the author used the company's Web site to identify library systems that fit that criterion. The author identified fifteen companies and then e-mailed each company explaining the project to find a contact person who had expertise in Unicode. The survey was then e-mailed to that contact person.

Results

Eleven of the fifteen companies responded to the survey. Not all of the companies gave permission for their name to be used in the article, so data is presented in numerical form.

Of the eleven companies that responded to the survey, six responded that their company had implemented the Unicode standard within their library system. Five had not (see table 1). Most of the companies that had already implemented Unicode were headquartered in other countries. Of the five companies that had not implemented the Unicode standard, four of them had plans to implement by the end of 2003. Only one company did not plan to implement Unicode, mentioning that in their market other character sets were more prevalent.

There are two common methods of implementing Unicode within the library system software. It can be implemented as the native character set or it can be enabled by mapping the characters in other character sets, such as EACC, to Unicode. Five companies responded that they had implemented Unicode natively, storing at least some of the data in Unicode (see table 2). When asked the open-ended question about what sort of data was stored in Unicode, three companies responded that all of the data was stored in Unicode (see table 3). Another company stores all records, queries, results, and configuration files in Unicode. One company responded that the main bibliographic record is stored in formal MARC structure but can optionally be stored in UTF-8, an encoding format that will be explained in more detail later in this article. The sixth company has enabled Unicode by mapping their own proprietary character sets to Unicode, although Unicode is the native character set of their software application which uses Java.

Five of the six library systems used Unicode for record editing (see table 4). Many library system companies had developed ways to display and input native scripts before Unicode came along by relying on special terminals or software. This is no longer necessary because library systems can take advantage of the input method editors for various languages incorporated into operating systems that support Unicode, such as Windows.

Six systems support the UTF-8 encoding form (see table 5), the only encoding form currently supported in the MARC 21 Specifications for Record Structure, Character Sets, and Exchange Media. Most Web browsers support UTF-8, and since online public access catalogs are displayed using Web browsers, it makes sense that library systems need to support that transformation format. However, as the software industry moves towards Unicode implementation around the world, support for more than just UTF-8 may be necessary in the future for easy exchange of records. For example, if you are trying to import records from a source that uses UTF-16 and your system supports UTF-8 only, then development will be required to make that exchange. Three systems also

Table 1. Unicode Implementation

	Yes	No
Has your company implemented the Unicode standard in any way within your library system?	6	5
Does your company plan to implement Unicode within the library system in the future?	4	1

Table 2. Native or Enabled Unicode

	Natively	Enabled
Does the library system use Unicode natively (i.e. is it built into the system) or has your library system enabled Unicode via some other means (e.g., mapping of other characters to Unicode)?	5	1

Table 3. Storing Data in Unicode

	Yes	No
Does the library system store any data in Unicode?	5	1

support UTF-16. Two systems support UTF-32 as well as the other encoding formats. UTF-16, with its use of two 16-bit code units, allows for the encoding of over a million characters. Thus, it can incorporate rare characters and many more of the East Asian ideographs than the 65,000 characters in 16-bit Unicode.

Sorting and indexing can be tricky issues. The Unicode standard provides a Unicode collation algorithm that specifies an ordering for all of the characters. However, for some languages this may not always produce the results that users expect. More development is required on the part of library system companies to produce results that make sense to a user. The survey question about sorting only addresses whether the library system uses Unicode to sort, not the ins and outs of how the system sorts for specific languages to meet user expectations (see table 6). For example, Chinese, Japanese, and Korean all share certain Chinese characters in common in their written scripts. A common way to sort the Chinese characters is by the number of strokes in a character. However, Japanese and Korean scripts contain Japanese and Korean characters as well as the Chinese characters. Japanese users expect sorting by the pronunciation of the Japanese characters in gojuon order. For Korean, the Chinese characters should sort by stroke, but the Korean characters should sort by Hangul (Korean alphabetic order). When speaking with companies about Unicode, it is important to understand how they deal with sorting of specific languages.

The survey also asked the companies which scripts their system supported. Each of the companies that had implemented Unicode responded to this question with a list of scripts such as Latin 1 and Arabic that were either supported, under development, or planned for the future. Most companies do not plan to support all of the scripts in the world, but are concentrating on the ones that their customers need. If you are a customer of a company that has implemented Unicode and want to take advantage of this capability, it is important for you to know which scripts are currently supported or will be supported in the future.

The last two questions on the survey inquired about the reasons companies decided to implement or not implement the Unicode standard. Companies in this extremely specialized and competitive area of business have to respond to global markets. Most answers centered around market forces that are driving the implementation, either responding to existing customer needs or expanding their business to other countries where populations use a variety of languages. One company emphasized that the increase in Chinese-speaking peoples in Australia influenced their decision to move to Unicode. Similar population migrations are happening around the globe, such as increasing populations of Spanish-speaking and Asian peoples in the United States. Companies respond to growth of multiple languages within their own country to better serve their existing customers as well as market to libraries in other countries. Another company mentioned that as soon as fonts and Web browsers that supported Unicode became available, it became feasible to provide multiple language support on one client PC, instead of using special terminals and software. Yet another company moved toward Unicode compliance because of their experience with foreign language publishing products that use SGML and XML, which use the Unicode character set. At first, they mapped their existing character sets to Unicode but eventually moved to a full native implementation. As more businesses and products begin to comply with the Unicode standard, compatibility for record exchange will accelerate the rate of Unicode implementation.

Table 4. Importing and Exporting Data in Unicode

	Yes	No	No Response
Does your library system use Unicode for record editing purposes?	5	1	--
Can your library system import data in Unicode?	5	--	1
Can your library system export data in Unicode?	4	1	1

Table 5. Transformation Formats

	UTF-8	UTF-16	UTF-32
Which transformation formats does your library system support?	6	3	2

Table 6. Sorting, Indexing, and Searching Using Unicode

	Yes	No
Does the library system use Unicode to sort any scripts?	3	3
Does the library system use Unicode to index any scripts?	5	1
Does the library system use Unicode to provide searching on any scripts?	6	0

Conclusion

Library system vendors have taken an active interest in Unicode and, with the exception of one company, all of the companies responding to the survey had either implemented Unicode or had plans to implement Unicode by the end of 2003. A number of factors have come together in recent years to contribute to the ease with which a company and a library can take advantage of Unicode. MARC 21 specifications now include support for the use of Universal Coded Character Set (ISO/IEC 10646) for exchange of MARC 21 records. Mappings to Unicode from MARC character sets have been completed. Web browsers now include support for Unicode. Unicode fonts that include many more characters are now available. The newer operating systems include support for Unicode, and some include input method editors, which allow the user to input characters in the native scripts of various languages.

However, companies are at different stages of development. Key questions for customers of companies that have implemented or plan to implement the standard are centered around which scripts have been implemented. Most library system vendors are implementing Unicode on a language-by-language basis to meet customer needs. Libraries will want to question vendors carefully about which scripts are supported. For example, vendors may implement support for Arabic, but this doesn't necessarily mean they include support for the additional extended Arabic characters used in such languages as Persian. Another key question would be how sorting of particular languages is handled. Because sorting can be tricky depending on the language, it would be beneficial to be able to examine another customer's site that is using Unicode to see how the characters sort. If the library needs to import and export records in Unicode, they must ask their company if that is possible and which encoding formats are supported.

Although it has been more than a decade since the Unicode standard was first developed, Unicode implementation has only been gaining momentum in recent years. Many factors have come together to make it easier to take advantage of Unicode. For companies that produce library systems, these and other forces are moving this process along. Competition in this small industry has pushed companies to market their systems in other countries. To compete in global markets, companies must be able to support multiple languages in their system. Companies have also responded to customer needs in their own country, where population migrations have caused other languages to become prevalent. Although by itself Unicode is not a complete solution for language support in library systems, it is certainly a foundation on which to build and will make development for multiple languages much easier.

Acknowledgements

Gary L. Smith, Senior Consulting Systems Analyst at OCLC, for providing feedback on the survey questions. Joan Aliprand, Senior Analyst at The Research Library Group (RLG), Inc., for providing information about the current state of Unicode at RLG. Unicode is a registered trademark of Unicode, Inc.. Eureka is a registered trademarks of RLG.

References and Notes

1. James Kass, Does Your Browser Support Multi-language? Accessed May 15, 2002, http://home.att.net/~jameskass.

2. Janet C. Erickson, "Options for Presentation of Multilingual Text: Use of the Unicode Standard," *Library Hi Tech* 15, no. 3/4 (1997): 172–88.

3. Foster J. Zhang and Marcia Lei Zeng, "Multiscript Information Processing on Crossroads: Demands for Shifting from Diverse Character Code Sets to the Unicode Standard in Library Applications," *IFLA Journal* 25, no. 3 (1999): 162–67.

4. Joan Aliprand, "The Unicode Standard: Its Scope, Design Principles, and Prospects for International Cataloging," *Library Resources and Technical Services* 44, no. 3 (July 2000): 160–67.

5. Vinod Chachra, Unicode and the World's Languages, paper presented at the ALA Annual Meeting, San Francisco, June 16, 2001.

6. Zhang and Zeng, "Multiscript Information Processing on Crossroads."

7. Ibid.

8. *Library Systems Newsletter* (Chicago: Library Technology Reports, American Library Association). The newsletter usually devotes the March issue to their annual survey of library system vendors.

9. Marshall Breeding, Library Technology Guides: Key Resources and Content Related to Library Automation. Accessed May 15, 2002, http://staffweb.library.vanderbilt.edu/breeding/ltg.html.

A Resource Description Device Used for More Efficient Library Services

Markos Dendrinos and Stelios Bakamidis

A special portable device designed for retrieving concise library resource descriptions by the user is presented in this article. The device reads a bar code attached to the back of the resource and searches for the corresponding information stored in a nonvolatile, rewritable memory. The information retrieved by the device can also be used in the loaning process. Resource descriptions can be simultaneously visual and dictated through a speech synthesizer, constituting a valuable tool for individuals with special needs, including the deaf and blind.

This article discusses a device designed to provide a concise description of each resource located in a certain section of a library. Description information is retrieved through the reading of a unique bar code attached to the back of each resource. This information can also be used for automating the loaning process, since the elements of the resource are already recorded and the only information to be added are the elements of the user and the loan and delivery dates. Resource descriptions can be simultaneously visual and dictated through a speech synthesizer, constituting a valuable tool for individuals with special needs, such as the deaf or blind.

What must be stressed here is the great importance of such a system for visually impaired persons. IFLA's Section of Libraries for the Blind (SLB) was established in 1983 as a forum for libraries for the blind. SLB participates in the annual IFLA conference and also in a biannual IFLA preconference for the section. The 2001 preconference, which took place in Washington, D.C., focused on increased information choices through Web-based technologies, future library services for blind students, digital delivery for the blind, and mainstreaming library services for blind and print-disabled users.[1]

Despite growing technological developments in the information and communications area, only a small percentage of documents are actually made available to the blind in accessible formats, which include speech output, braille output, tactile devices, or even simple adjustments to a browser.[2]

Integration of blind and visually impaired persons into schools, universities, and training centers is being considered through projects such as BrailleNet.[3] BrailleNet concerns Internet document delivery and aims to achieve integration through Web-accessible assistive technologies and teaching materials. The delivery of these special books is further enabled through cooperation with publishers, adaptation centers, and printing centers.

National Library Service for the Blind and Physically Handicapped (NLS), Library of Congress, is making use of the Internet to deliver a number of its services.[4] A continuously growing number of Web-Braille titles (currently 3,800 titles) has been made available to 1,500 users.

Speech input/output through speech recognition/synthesis is the most user-friendly interface for offering information in the case of visual inefficiencies, contributing greatly to a more efficient interoperability of the previously mentioned initiatives.

Presentation of the Operations of the Resource Description Device

The library is organized in a series of thematic sections. Each section consists of a number of resources, such as books, serials, images, videos, audio, and maps. Each resource is described through certain bibliographic fields along with a summary text. The whole description is stored in a special bibliographic database, allowing any field to be retrieved through tools included in related management software packages. Each library section is to be related to a certain device. This infrared device will be connected to the computer server with the mentioned database. In this way the resource descriptions corresponding to a section can be downloaded and stored in the device's nonvolatile, rewritable memory. Obviously, these data can be changed or deleted, following relative modifications in the real world state.

The library user can get the device from the information desk and go to the part of the section he or she is interested in. As the device is passed over the back of each resource, the unique bar code attached to the resource will trigger the device's sensor, which will retrieve the relative part of the memory containing the resource description. The retrieved information will be displayed in a low-energy-consumption LCD display and at the same time dictated through a speech synthesizer. The visual part is of great importance for persons with acoustic problems, while the acoustic part is greatly sup-

Markos Dendrinos (mdendr@teiath.gr) is an Assistant Professor in the Department of Library Studies in the Technological Educational Institution of Athens (TEI-A), Greece, and a Researcher in the Speech Technology Department of the Institute for Language and Speech Processing (ILSP), Athens, Greece. **Stelios Bakamidis** (bakam@ilsp.gr) is the Head of the Speech Technology Department of the Institute for Language and Speech Processing (ILSP), Athens, Greece.

portive for persons with visual problems (see figure 1).

Speech synthesis preprocess includes an automatic language detection and routing module followed by a multilingual text-to-speech (TTS) module. TTS is based on specific rules dependent on the certain language. The system is to operate with all the official European Union (EU) languages. The inclusion of Greek in such a system is of national strategic importance, aiming at the equal linguistic participation of EU partners independently of the corresponding populations. The speech synthesizer is described below.

Lastly, the device is to be equipped with all necessary navigation, control, and scroll buttons as well as an acoustic and visual help guide, offering an additive value in the cases of blind and deaf people.

After the user is informed about the resource, he can get the resource to read visually or acoustically (through ear phones) in an area of the library. Another alternative is to loan the resource. All the information already retrieved by the description device can be integrated with the elements of the borrower, the loan date, and the delivery date, thus creating a fulfilling of each loan action.

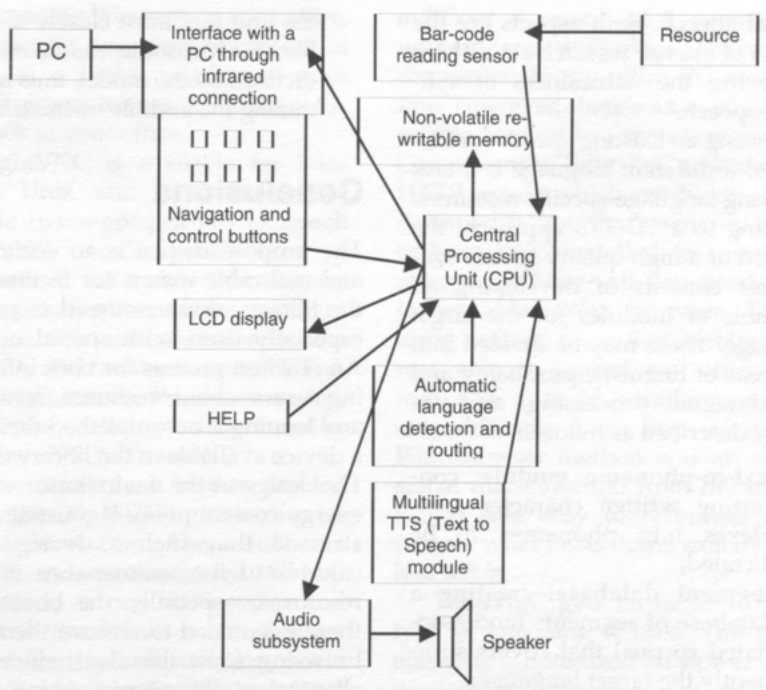


Figure 1. Structure and Operations of the Resource Description Device

The Speech Synthesizer

A number of commercial and laboratory prototype systems have been presented for text-to-speech (TTS) synthesis. The majority of them are based on one of the three most popular paradigms:

- Rule-based speech synthesis.[5]
- Speech synthesis based on time-domain techniques.[6]
- Speech synthesis based on articulatory models of the human speech production system.

Each method possesses quite different characteristics that renders it more suitable for specific application areas. Where speed of execution is most important, a time-domain technique is the prime candidate. For memory-sensitive application environments, formant-based techniques present a distinct advantage.

Modeling the human speech production system is a demanding task, since the incorporated articulatory models require intense calculations. This fact severely inhibits the implementation of articulatory models into real-world commercial applications.

Time-domain text-to-speech (TD-TTS) conversion relies on a large database of prerecorded natural speech segments that are appropriately concatenated to obtain the speech transcription of arbitrary text.[7] By employing sophisticated algorithms for seaming the discrete segments, one can achieve quite natural synthetic speech.[8] Rule-based text-to-speech conversion (RB-TTS), on the other hand, models the human speech production system more closely, requiring a more profound examination and a direct modeling of all the phenomena involved. A number of high-quality, state-of-the-art systems based on RB-TTS have been presented confirming the value of this method.

Synthetic speech quality, especially naturalness, is largely dependent on the sophistication of the prosodic modeling and the prosodic rules employed. On the other hand, detailed prosodic implementation increases substantially the intelligibility of the system even at segmental levels.

The majority of TTS systems are based on sentence-level prosody, which provides various degrees of intelligibility but hardly any quasi-natural output from a prosodic and thus phonetic point of view. The main directions for improving the naturalness of synthetic speech involve studying the synthetic signal quality as well as the prosodic modeling of

natural speech. Both aspects are the subject of intense research activity for improving the naturalness of synthetic speech.[9]

Porting an existing speech synthesizer to a different language is a task requiring language-specific resources. Focusing to a TD-TTS approach, the creation of a high-quality speech synthesizer consists of developing an ensemble of modules for the target language. These may be divided into the areas of linguistic processing and digital signal processing, and are briefly described as follows:

- **text-to-phoneme module:** converting written character complexes into phonemes to be dictated;
- **segment database:** creating a database of segments (and associated corpus) that covers sufficiently the target language;
- **text decomposer**: deriving an algorithm for decomposing text into segments;
- **prosodic modeling:** creating a prosody generator for the target language that provides the desired synthetic speech quality;
- **speech corpus:** obtaining an adequate corpus of prerecorded utterances, which will provide the basis for defining speech segments in various environments to be concatenated during synthesis;
- **synthesis algorithms:** designing the algorithms that join the segments so as to generate the synthetic speech signal.
- **unit selection:** providing multiple instances of each segment possessing different prosodic properties in the database to improve the speech quality. An algorithm is then used to select the unit that most closely resembles the prosodic characteristics dictated by the model, thus minimizing the audible mismatches.[10]

Conclusions

The propose device is an economic and realizable means for facilitating the library services offered to users, especially users with special needs. An efficient process for both informing users about resource contents and loaning is accomplished through a device available in the library desk. The design of the device ensures low energy consumption. It must also be stressed that such a device contributes to the maintenance of the resources, especially the books, as there is no need to remove them for browsing from the shelf, since the client takes all the necessary information from the bar code attached to the back of the resource.

References

1. Jenny Craven, The Development of Digital Libraries for Blind and Visually Impaired People. Accessed Feb. 27, 2002, www.ariadne.ac.uk/issue30/ifla.

2. K. Miesenberg, "Future Library Services: Developing Research Skills among Blind Students," in *Digital Libraries of the Blind and the Culture Of Learning in the Information Age: Conference Proceedings of the IFLA SLB Preconference* (Washington, D.C.: IFLA/SLB, 2001).

3. D. Burger, "BrailleNet: Digital Document Delivery for the Blind in France," in *Digital Libraries of the Blind and the Culture Of Learning in the Information Age: Conference Proceedings of the IFLA SLB Preconference* (Washington, D.C.: IFLA/SLB, 2001).

4. C. Sung, "The Future of Lifelong Learning in the Next Generation of Library Services," in *Digital Libraries of the Blind and the Culture Of Learning in the Information Age: Conference Proceedings of the IFLA SLB Preconference* (Washington, D.C.: IFLA/SLB, 2001).

5. A. Conkie and S. Isard, "Optimal Coupling of Diphones," in J. Van Santen et al., eds., *Progress in Speech Synthesis* (New York: Springer-Verlag, 1997), 279–82.

6. Conkie and Isard, "Optimal Coupling of Diphones"; E. Moulines and F. Charpentier, "Pitch Synchronous Waveform Processing Techniques for Text-to-Speech Using Diphones," *Speech Communication* 9, no. 5–6 (autumn 1993): 453–70.

7. T. Dutoit, *An Introduction to Text-to-Speech Synthesis* (Dordrecht, Netherlands: Kluwer Academic Pr., 1997).

8. T. Dutoit and H. Leich, "Text-to-Speech Synthesis Based on a MBE Resynthesis," *Speech Communication* 13, no. 3–4 (autumn 1993): 435–40; Y. Stylianou, *Harmonic Plus Noise Models for Speech, Combined with Statistical Methods, for Speech and Speaker Modification*, Ph.D. thesis, École Nationale Supérieure des Télécommunications, 1996; Y. Stylianou, "Removing Linear Phase Mismatches in Concatenative Speech Synthesis," *IEEE Transactions on Speech and Audio Processing* 9, no. 3 (Mar. 2001): 232–39.

9. E. Keller et al., *Improvements in Speech Synthesis: COST 258: The Naturalness of Synthetic Speech* (Chichester, England: John Wiley & Sons, 2002).

10. Conkie and Isard, "Optimal Coupling of Diphones"; M. Founda et al., *Proceedings of the Eurospeech—2001 Conference* vol. 2 (Aalborg, Denmark: Center for PersonKommunikation, 2001), 837–40; E. A. M. Klabbers and R. Veldhuis, "On the Reduction of Concatenation Artifacts in Diphone Synthesis," *Proceedings of the ICSLP 98 Conference*, vol. 5 (Sidney, Australia: ASSTA, 1998), 1,983–86.

Software Reviews

Andy Boze, Editor

Since this column will appear in the December issue, following is a trio of useful applications suitable for holiday gifts for yourself and others. They all have the advantage of being free or inexpensive.

TightVNC is a popular remote control program that makes it possible to use a computer at another location as though you are sitting in front of it. WinSCP is a file transfer application that uses the Secure Copy Protocol. All data are encrypted so that you don't have to worry about your files and passwords passing over the network in clear text and being vulnerable to snooping. Finally, Mailbag Assistant makes it easy to manage all that e-mail that you've got piling up on your hard drive.

TightVNC 1.2.6

Constantin Kaplinsky
E-mail: const@ce.cctpu.edu.ru
www.tightvnc.com

Price: free
System requirements: Windows 95, 98, NT, 2000, XP, or ME; Unix/Linux (various versions)

You're at home, or maybe you're at a conference. Wherever you are, you'd like to use that computer back in your office where your files and your software are. With TightVNC you can remotely control your office PC, or any other PC, from wherever you happen to be, as long as you have a computer handy.

TightVNC is one of a family of products, all based upon Virtual Network Computing (VNC), which was developed at AT&T Laboratories in Cambridge, England. To quote from AT&T's Web site, VNC is "a remote display system which allows you to view a computing 'desktop' environment not only on the machine where it is running, but from anywhere on the Internet and from a wide variety of machine architectures." What this means is that you can have a PC running Windows in your office, and you can use your Macintosh at home or a Linux machine at your friend's house to view and control the PC back in your office.

TightVNC is available for Windows, Unix, and Java. If you use Google (www.google.com) to search for "VNC," you'll quickly turn up members of the family that support other operating systems. Interestingly, there are even versions of VNC for the Palm OS and for the Pocket PC. This review refers to the Windows version, which was used on PCs running Windows NT, 2000, and XP.

The installer can be downloaded from the TightVNC Web site at www.tightvnc.com. The installation is quick and uncomplicated. All you need to do is run the installer and accept all the default values that are offered. During the installation, a new folder is added to the Windows Start menu. One of the items in the folder is for installing TightVNC as a service. You'll want to do that if the machine is the one you want to control remotely. Installing the TightVNC service means that it is automatically loaded at every system start. That's useful in case the computer gets restarted while you are away, otherwise you wouldn't be able to connect to it until someone logs in and manually starts the TightVNC server. Of course, in order to control a computer remotely, you need to have VNC installed on a second one.

I use TightVNC on my office and home PCs. Usually I want to connect to my office PC from home, so I keep the VNC server running on the office PC and use the viewer on the home PC to make the connection. I'm fortunate enough to have a cable modem at home, so the PC there is online all the time, just like the office PC. I keep the VNC server running on my home machine in case I want to connect to it from the office.

You might wonder how you can use VNC from a machine that you don't own, if you're at a conference, for example. I installed TightVNC on a PC at a recent ALA conference without any difficulties. Making a connection to my office PC worked fine. However, there's an even easier way to connect to a remote machine. Conveniently, TightVNC includes an HTTP server, which can be disabled if you wish. You don't actually need to have VNC installed to control a remote machine—all you need is a Java-enabled Web browser. Thus, using nothing more than Netscape, I was able to control my office PC from a PC at an ALA conference. The only difficulty I ran into using the Web browser method was an occasional disconnection from my office PC. It was easy to reconnect and resume what I was doing exactly as I had left it.

TightVNC gets its name from a type of encoding it uses. The tight encoding is optimized for slower connections, making TightVNC a good choice for use with dial-up modem connections. You don't have to use tight encoding. In fact, you can choose from among several types of encoding and compression methods. It's worth experimenting with different settings to tweak performance and image quality. By default, tight encoding uses a high degree of graphics compression so that less data has to pass over the network. I've found that high compression produces a perfectly acceptable image quality. If you demand that colors and clarity be exactly as on the remote monitor, you can use lower compression or disable compression, at the expense of performance.

Using TightVNC to connect to a remote machine is simple. From the Windows Start menu, you select TightVNC Viewer, and fill in the dialog box with the host name or IP address of the remote machine. You can also select specific encoding and compression settings and various other options. Then click the OK button to make the connection. The remote machine will prompt you for a password that you should have set after installing the software on the

remote machine. Without setting a password, anybody could connect to your machine and use it. Depending upon network speed and the options you select, you'll see the remote screen almost instantaneously or after a few seconds. At this point, you can use the remote machine exactly as though you were sitting directly in front of it.

If you frequently connect to the same machine using the same settings, you can save all the connection information. While the connection is active, you right click on the TightVNC button on the Windows Task Bar. One of the options on the pop-up menu is to save the connection information. Selecting it will save all the information to a file. Subsequently, you can double click on the file to start a new connection using all the saved settings.

TightVNC has a couple of features that I find particularly convenient. The screen area of my office PC is set at 1600x1200 pixels, but my home PC is set at 1280x1024 pixels. When I connect from home to office, the remote screen is too large to fit entirely on the local screen. Moving the mouse cursor to any edge of the screen scrolls the view in that direction, so I don't have to use scroll bars or press keys. What I like even better is that I can scale the remote screen so that it will fit entirely on my local screen. This feature is labeled as experimental, but it has worked well for me. I used a scaling of three-quarters and was surprised that the text on the remote screen was still very legible. Also handy is the fact that anything I copy to the clipboard of the remote machine can be pasted into a local document.

One thing that is slightly inconvenient about TightVNC is the inability to use the local keyboard to send a Ctrl-Alt-Del to the remote machine. To do that, you have to right click on the TightVNC button on the Windows Task Bar and then select Send Ctl-Alt-Del from the pop-up menu. A feature that I would like to see added to TightVNC is to be able to access the remote machine's file system so that I can copy files directly between the remote and local machines. For now, you can always connect to the remote machine and transfer any files to an intermediate location that both machines can access, using FTP, for example. A toolbar for common functions would also be useful, and apparently is planned for a future release.

It's important to remember that installing software of this type can expose your computer to security risks. Talk to your IT or network support staff to be sure that it is installed and configured properly.

WinSCP 2.0 Beta

Martin Prikryl
martin.prikryl@seznam.cz
http://winscp.vse.cz/

Price: free

System requirements: Windows 95, 98, NT, 2000, XP, or ME; SSH server for remote connection

Organizations are becoming increasingly conscious of security. Instead of Telnet and FTP, Secure SHell (SSH) is becoming increasingly common. If you want to transfer files, this can be something of a problem since Secure CoPy (SCP) clients, which use SSH, aren't as numerous as FTP clients. Most Web browsers will let you download files by FTP, and Netscape Communicator and Internet Explorer will even let you upload files by FTP. But what if your files are on a secure server? Your organization probably provides your work computer with an SSH client, but if you are at home or elsewhere you'll need an SCP client to transfer files. WinSCP is just the thing.

You can download WinSCP from the author's Web site at http://winscp.vse.cz. No installation is necessary. There's a single executable file that you can run immediately after downloading. An older version does have an installer, if you are so inclined.

After starting WinSCP, you'll get a dialog box with several tabs for various options (see figure 1). Since the default options worked when I connected to my server, I didn't change most of them. In the Basic tab, you fill in the host name you want to connect to, your user name, and your password. You can select from one of two interface styles. One style is similar to the Norton Commander with two panels, one for local files and the other for remote files. The other interface choice resembles the single-panel Windows Explorer, showing only the remote files (see figure 2). If you plan on using this connection again, you can store the session settings for future use.

WinSCP has a drag-and-drop interface, so you can use it almost exactly as you would Windows Explorer. Dragging any file from the remote panel to a local panel will copy the file locally. Dragging a local file to the remote panel will copy it to the server. Right clicking on a file will let you copy, move, rename, or delete it. You can even have two or more WinSCP sessions running simultaneously to different servers and copy files between them.

WinSCP support SSH1 and SSH2, and authentication can be done either

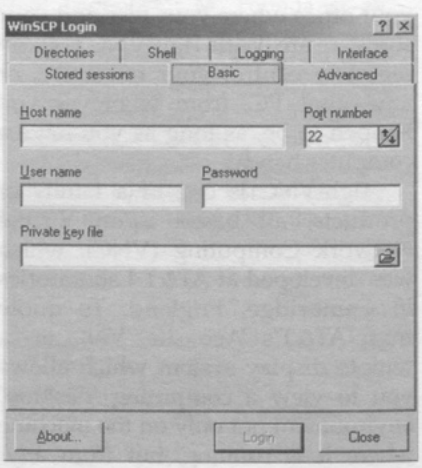


Figure 1. WinSCP Login Dialog Box

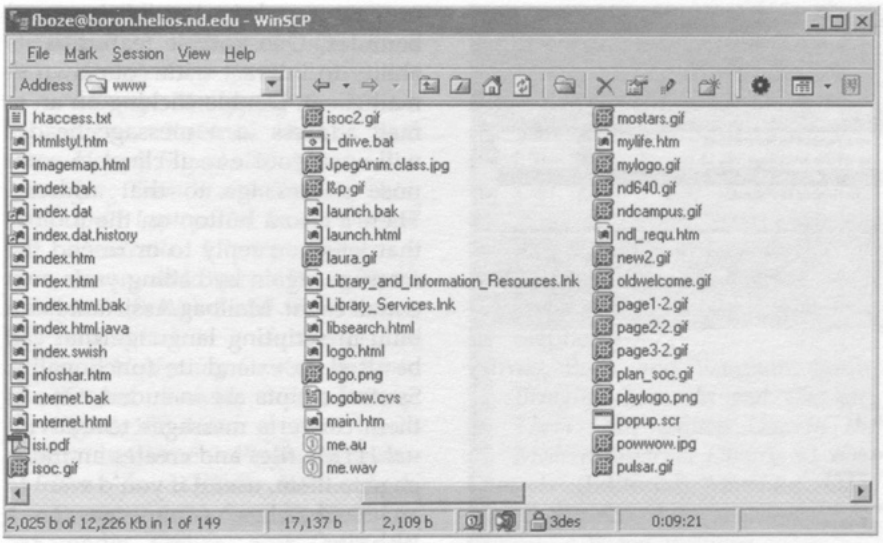


Figure 2. WinSCP Connected to Server Using Windows Explorer Style

by using a user name and password or by using an RSA key. The author's Web site includes a utility that you can use to generate an RSA key file. The program supports DES, 3DES, and Blowfish encryption and also support compression.

Even though WinSCP 2.0 is advertised as being a beta version, it has worked without any problems for me. For those uncomfortable using beta software, version 1.0 is also available from the Web site.

Mailbag Assistant 3.0

Fookes Software
Av. Eugène-Pittard 22 Ter
CH - 1206 Geneva
Switzerland
+41/22-789 58 44
E-mail: sales@fookes.com
www.fookes.com

Price: $29.95

System requirements: Windows 9x, 2000, NT4, Me, or XP operating system; 16 MB of RAM (32 MB for NT), 32 MB recommended; approximately 4 MB of free disk space, 8 MB recommended; Intel® 486 or greater processor

Do you get a lot of e-mail, and worse, do you never delete any of it? If you've saved every e-mail you have ever received, or if you just need to manipulate your e-mail efficiently, Mailbag Assistant may be exactly the thing you need.

Mailbag Assistant can open mailboxes (e-mail files) from about a dozen different programs, including Agent, Calypso, Eudora, Foxmail, Netscape, Netscape 6, Outlook Express, Pegasus, and more. Oddly, perhaps, it can't open Outlook mailboxes, but the help file does mention that Outlook Express can import Outlook mailboxes and then Mailbag Assistant can open the imported mailboxes. You can open one or many mailboxes simultaneously, even if they are from different programs. Mailbag Assistant doesn't change anything in the original mailboxes, so there's no danger that your e-mail program will no longer work with them.

The first time you start Mailbag Assistant, you'll get a wizard to help you select your mailboxes. The first page of the wizard asks you what type of mailbox you want to open. The next page asks you to locate the mailboxes and whether you want to open all of them or just specific ones. The third page of the wizard lets you specify search criteria if you want to open only messages containing particular text or dates. The final page of the wizard asks whether you want to perform any specific tasks with the messages that will be opened. You can opt to do nothing, to sort the messages into groups, to make an archive, or to export them into various formats. If you choose to do nothing, a list of the messages is displayed in a view probably much like your own e-mail program uses. The display uses two panels, the upper one for the list of messages and the lower one to show the selected message (see figure 3).

The upper panel actually consists of several tabbed views, one of which, Grid View-Main, is the list of messages. The Message Header view lets you see the raw message header, and the Message Body view lets you see just the body of the message. These views can be handy if you want, for example, to clip just certain parts of many messages. If you right click in the Message Body view and select Copy to Output, the message body will be copied into a new tabbed view labeled Output View, which acts as a clipboard. Each time you select Copy to Output, the new text is appended to what is already there. Ultimately, you can copy whatever you have in the Output View into another application or you can print it.

Within Grid View-Main there are several ways to manipulate the messages. As you review the messages, you can transfer any of them to a subset view to keep them separate from the main list. A useful option is the ability to extract attachments from messages and save them separately. From the menu you can also extract HTML pages and lists of e-mail

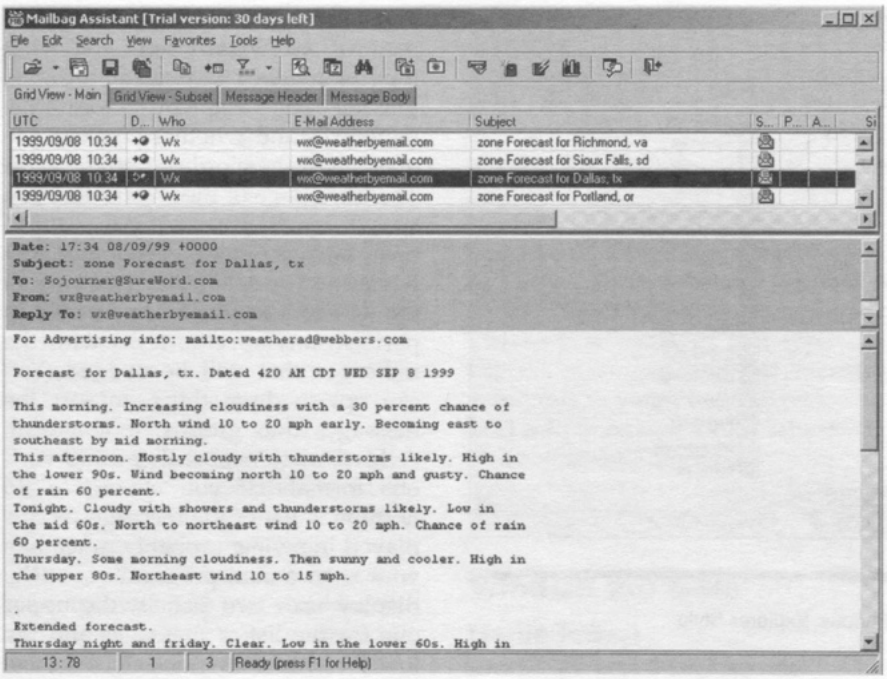


Figure 3. Mailbag Main Display with Two Panels

addresses from messages. Messages can be sorted in several ways, including by subject, domain (e.g., yahoo.com), or mailbox name.

Mailbag Assistant has powerful capabilities for searching and filtering messages, including Boolean operators and options such as date ranges, regular expressions, and Soundex. One notable feature is its ability to interact with your own e-mail client. Double clicking on an e-mail address in a message header will open your e-mail client to compose a message to that address. There's also a button on the toolbar that let's you reply to or resend the message, again by calling your own e-mail client. Mailbag Assistant has a built-in scripting language that can be used to extend its functionality. Several scripts are included. One of them converts messages to individual HTML files and creates an index page to them, useful if you'd want to archive batches of messages on a Web site.

Mailbag Assistant is rather like a Swiss army knife for e-mail management. There's something here for everyone and a there is a tool for just about any conceivable task. As much as it can do, I found that it was easy to use. After a little tinkering, I quickly became comfortable with it, although learning its advanced features, such as scripting, will take a little time and practice. The program can be downloaded from the company's Web site for a thirty-day trial at www.fookes.com/mailbag.

Index to Advertisers

Information Technology and Libraries
Index to Volume 21 (2002)

Compiled by Charles Anderson

STATEMENT OF OWNERSHIP, MANAGEMENT, AND CIRCULATION

Information Technology and Libraries, Publication No. 280-800, is published quarterly in March, June, September, and December by the Library Information and Technology Association, American Library Association, 50 E. Huron St., Chicago, Illinois 60611-2795. Editor: Dan Marmion, 221 Hesburgh Library, Notre Dame, IN 46556-5629. Annual subscription price, $30. Printed in U.S.A. with periodical-class postage paid at Chicago, Illinois, and other locations. As a nonprofit organization authorized to mail at special rates (DMM Section 424.12 only), the purpose, function, and nonprofit status for federal income tax purposes have not changed during the preceding twelve months.

EXTENT AND NATURE OF CIRCULATION

(Average figures denote the average number of copies printed each issue during the preceding twelve months; actual figures denote actual number of copies of single issue published nearest to filing date: June 2002 issue). Total number of copies printed: average, 6,195; actual 6,300. Sales through dealers and carriers, street vendors, and counter sales: none. Paid or requested mail subscriptions: average, 5,913; actual, 5,740. Free distribution (total): average, 50; actual, 48. Total distribution: average, 5,963; actual 5,788. Office use, leftover, unaccounted, spoiled after printing: average, 232; actual, 512. Total: average, 6,195; actual, 6,300. Percentage paid: average, 99.16; actual, 99.17.

Statement of Ownership, Management, and Circulation (PS Form 3526, October 1999) filed with the United States Post Office Postmaster in Chicago, October 1, 2002.

5th ANNUAL

LITA NATIONAL FORUM

SPONSORS

AWE Alternative Work Environments
www.awelynx.com

Axonix Corporation
www.axonix.com

Dell Computer Corporation
www.dell.com

Fretwell-Downing, Inc.
www.fdusa.com

Gaylord
www.gaylord.com

Innovative Interfaces, Inc.
www.iii.com

LSSI
www.vrtoolkit.net

OCLC
www.oclc.org

ProQuest
www.il.proquest.com

Swets Blackwell
www.swetsblackwell.com

TechBooks
www.techbooks.com

Veicon Technology, Inc.
www.veicon.com

THANK YOU!